



City Research Online

City, University of London Institutional Repository

Citation: Goodwin, S. (2015). Visualisation for household energy analysis: techniques for exploring multiple variables across scale and geography. (Unpublished Doctoral thesis, City University London)

This is the draft version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/14535/>

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**VISUALISATION FOR
HOUSEHOLD ENERGY ANALYSIS:**
TECHNIQUES FOR EXPLORING MULTIPLE
VARIABLES ACROSS SCALE AND GEOGRAPHY

SARAH M. GOODWIN

A THESIS SUBMITTED FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
IN GEOGRAPHICAL INFORMATION SCIENCE

AUGUST 2015



**CITY UNIVERSITY
LONDON**

GiCENTRE, DEPARTMENT OF COMPUTER SCIENCE



THE FOLLOWING PAPER HAS PREVIOUSLY BEEN PUBLISHED:

pp 218-226:

Goodwin, S., Dykes, J., Jones, S., Dillingham, I., Dove, G., Duffy, A., Kachkaev, A., Slingsby, A., and Wood, J., Creative user-centered visualization design for energy analysts and modelers. In *IEEE Transactions on Visualization and Computer Graphics*, 19(12), pp. 2516-2525.

doi: [10.1109/TVCG.2013.145](https://doi.org/10.1109/TVCG.2013.145)

(c) 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

THE FOLLOWING PAPER AND POSTER WERE PRESENTED AT A CONFERENCE:

pp 227-229:

Goodwin, S., Dykes, J. & Slingsby, A. (2014). *Visualizing the Effects of Scale and Geography in Multivariate Comparison*. Poster presented at the IEEE Conference on Visual Analytics Science and Technology, 09-11-2014 - 14-11-2014, Paris, France.

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. Also presented at Workshop on GeoVisual Analytics: Interactivity, Dynamics, and Scale at the Eighth International Conference on Geographic Information Science, 23-09-2014 - 26-09-2014, Vienna, Austria.

Contents

List of Tables	vii
List of Figures	ix
Acknowledgements	xiv
Publications	xvi
Declaration	xviii
Abstract	xix
Abbreviations	xx
1 Introduction	1
1.1 Research Context	1
1.2 Aims and Research Questions	5
1.3 Thesis Scope and Contributions	8
1.4 Research Methods	9
1.5 Use of Academic Literature	9
1.6 Research Ethics	10
1.7 Thesis Structure	11
2 Literature Review	13
2.1 Energy Consumption Analysis	13
2.1.1 Household Consumption Analysis	13
2.1.2 Smarter Energy Analysis	15
2.1.3 Energy Consumption Analysis Summary	17
2.2 Data Visualisation: Design, Application & Creativity	18
2.2.1 Visualisation	18
2.2.2 Visual Design	19
2.2.3 Energy Consumption Visualisation	21
2.2.4 Creativity and Visualisation	23
2.2.5 Data Visualisation Summary	24
2.3 Geodemographic Profiling	24

2.3.1	Geodemographics	25
2.3.2	Energy-based geodemographics	26
2.3.3	Creating Geodemographics	28
2.3.4	A Four Stage Process	38
2.3.5	Visualising Geodemographics	40
2.3.6	Geodemographics Summary	42
2.4	Multivariate Comparison	42
2.4.1	Distribution	44
2.4.2	Correlation	45
2.4.3	Geographical Variation	48
2.4.4	Variable Scale	51
2.4.5	Visual Parameter Space Analysis (vPSA)	51
2.4.6	Multivariate Comparison Summary	52
2.5	Chapter Summary	53
3	Creative Energy Domain Exploration	55
3.1	Smart Home Project Description	56
3.2	Creative Requirements Workshop	56
3.2.1	Wishful Thinking	58
3.2.2	Constraint Removal	59
3.2.3	Visualisation Awareness using Analogical Reasoning	59
3.2.4	Storyboarding	60
3.2.5	Requirements Workshop Outcomes	61
3.3	Smart Home Data Availability	62
3.4	Design Workshop, Development & Feedback	63
3.5	Evaluation: Prototype Designs, Creativity & Process	64
3.5.1	User Evaluation Session	64
3.5.2	The Four Prototypes	65
3.5.3	Evaluation with the Data Modellers	72
3.5.4	Creative Design Process	73
3.5.5	Evaluation Conclusions	74
3.6	The Who? and Where? of Energy Consumption	75
3.7	Chapter Summary	78
4	Exploring Energy-Based Geodemographics	79
4.1	Exploratory Visual Analysis	80
4.1.1	Data and Limitations	80
4.1.2	Visual Technique	81
4.1.3	Visual Analysis	82
4.1.4	Summary of Exploratory Visual Analysis	85
4.2	Exploring a Visual Geodemographic Process	85
4.2.1	Investigation of the Available Tools	85

4.2.2	Requirements for Visual Representations	87
4.2.3	Visualisation Examples and Ideas	88
4.2.4	Summary of a Visual Geodemographic Process	93
4.3	Candidate Energy Variable Options (for the UK)	93
4.3.1	Electricity and Gas Consumption	93
4.3.2	Other Fuel Consumption and Central Heating Types	95
4.3.3	Energy Ratings	96
4.3.4	Heat Loss and Insulation	96
4.3.5	Fuel Poverty	96
4.3.6	Household, Demographic and Socio-Economic Variables	97
4.3.7	Candidate Energy Variable Options Summary	98
4.4	Chapter Summary	98

5 Geography and Scale: Data Preparation 101

5.1	Geographical Variation: Local Statistics	101
5.1.1	Calculating Local Statistics	103
5.1.2	Locality Calculation Sensitivity	104
5.2	Dimensions of Scale	106
5.3	Spatial Scale	106
5.3.1	Scale in the Variable Selection Process	107
5.3.2	Scale Sensitivity in Variable Selection	108
5.3.3	Spatial Scale in Geodemographic Examples	110
5.4	Data Preparation	110
5.4.1	Input Scale	111
5.4.2	Standardised Scale	112
5.4.3	Locality Scale	113
5.4.4	Output Scale	114
5.5	Data Calculation Method	114
5.5.1	Describing the Method	115
5.5.2	Statistical and Structural Outputs	116
5.5.3	Four Types of Scale in the Four Stage Process	118
5.6	Chapter Summary	119

6 Building and Visualising the Framework 121

6.1	Building the Framework	122
6.2	Potential Visual Representations of the Framework	124
6.3	Visualising the Framework	127
6.3.1	Prototype Design Decisions	128
6.3.2	Additional Functionality	150
6.3.3	Development Techniques	151
6.3.4	Referencing the Framework	151
6.4	Chapter Summary	153

7	Validating the Prototype Design	155
7.1	Appropriateness of the Prototype Design	156
7.2	Visual Exploration of Energy Variables	157
7.2.1	Global Skewness	157
7.2.2	Global Correlation	158
7.2.3	Global Geographical Variation	159
7.2.4	Global Statistics across Scale	161
7.2.5	Adding Locality into Variable Selection	163
7.2.6	Summary of Visual Exploration of Energy Variables	165
7.3	Geo-visual Parameter Sensitivity Analysis	165
7.3.1	Sensitivity of Geography: Varying N	166
7.3.2	Sensitivity of Scale: Varying SR	172
7.3.3	Sensitivity of Transformation: Logged Scale	175
7.3.4	Summary of Parameter Sensitivity	175
7.4	Identifying Discriminating OAC Variables	177
7.5	OAC Expert Feedback	178
7.6	Visualisation for Geodemographic Variable Selection	180
7.7	Possible Extensions for the Prototype	181
7.8	Chapter Summary	184
8	Further Applicability of the Framework	185
8.1	Designs for Future Work	185
8.1.1	Brainstorming Hierarchical Designs	186
8.1.2	Multiple Scale Mosaics	187
8.1.3	Locality and Scale Mosaics	188
8.1.4	Designs for Future Work Summary	189
8.2	Dynamic Geodemographic noClassifications	189
8.3	Additional Scenarios	190
8.3.1	Scenario 2: Smart Home Analytics	191
8.3.2	Scenario 3: Survey Response Modelling	194
8.4	Academic Feedback	197
8.5	Chapter Summary	198
9	Discussion and Conclusions	199
9.1	Summary of Thesis	199
9.2	Scope of Research	200
9.3	Research Discussion	200
9.3.1	Creative Domain Exploration	200
9.3.2	Energy Profiling: static and dynamic approaches	202
9.3.3	Changing Channel from Variables to Process	203
9.3.4	Visualising Geodemographic Variable Selection	204
9.3.5	Exploring Sensitivities through PSA	206

9.3.6	Visualising Sensitivities through gvPSA	206
9.3.7	The Design Science Approach	207
9.4	Research Contributions	210
9.5	Research Limitations	211
9.6	Future Work	213
9.7	Final Conclusions	214
A	Academic Impact	216
A.1	Article: TVCG 2013, 19(12), pp.2516-2525	217
A.2	Short Paper: IEEE VIS 2014, Paris, FR	227
A.2.1	Poster: IEEE VIS 2014, Paris, FR	229
A.3	Short Paper: IEEE VIS 2012, Seattle, USA	230
A.3.1	Poster: IEEE VIS 2012, Seattle, USA	232
A.4	Short Paper: GISRUK 2012, Lancaster, UK	233
A.4.1	Poster: GISRUK 2012, Lancaster, UK, - Won ‘Best Poster’	238
A.5	Abstract: NACIS 2013, Greenville, USA	239
A.5.1	Presentation: NACIS 2013, Greenville, USA	240
A.6	Paper: PhD Symposium 2012, Birmingham, UK	243
B	Additional Material	246
B.1	Smart Home Project Consent Form	247
B.2	Photos from the Requirements Workshop	248
B.3	Aspirations and Storyboards from Workshop	249
B.4	Concept Map of Energy Analysis Possibilities	251
B.5	Photos from the Internal Design Workshop	252
B.6	Questionnaire of Appropriateness	254
B.7	Generating Geodemographics Requirements	258
B.8	167 variables considered for OAC 2011 from Gale, 2014,pp.224	259
B.9	Final 78 Variables used in Framework Prototype	261
B.10	Brainstorming Hierarchical Designs	262
	References	265

List of Tables

2.1	Differing priorities that guide geodemographic classification (from Singleton and Longley, 2009a, pp.293)	29
3.1	Aspirations revealed in ‘ <i>Know/Do/See</i> ’ and ‘ <i>What next?</i> ’. Numbers show total aspirations established and those deemed feasible. Table 1 from Goodwin et al. (2013, pp.2519)	58
3.2	Prototype Enhancements, Table 2 from Goodwin et al. (2013)	64
3.3	Evaluation Process, Table 3 from Goodwin et al. (2013)	64
3.4	Aspirations relating to the segmentation of data by population or lifestyle characteristics	78
4.1	User stories for visual and analytical possibilities for the process of generating geodemographics created from the know/do/see activity. 1-34 in bold are relevant to variable selection	89
4.2	Potential candidate variables from OAC 2001 & 2011 and additional Energy domain variables relating to central heating fuel types, electricity and gas consumption and a fuel poverty indicator	99
5.1	The five aggregations of SR used in the standardise stage (StR)	112
6.1	The layout of the framework with the names of the cells and the goals of the three rows representing number of locations where local statistics are calculated (L) and the four columns representing numbers of variables (V) .	123
6.2	The ability to make comparisons when visualising multiple scale resolutions, scale extents or multiple transformations with increasing numbers of variables (V) and local summaries (L)	125
6.3	Statistical (top row) and spatial (italics in bottom row) visual possibilities within the framework as V and L increase	126
6.4	The prototype illustrated within the framework with the represented statistical and geographical (italicised) visuals and panels; P1, P2 and P3. Blue arrows 1-5 highlight transitions discussed	153

7.1	A selection of variables to demonstrate the comparison of the visual components of skewness and correlation with gas and electricity consumption when N in locality is varied (25, 50 and 100) for LAD. Three colour schemes (see Legend in Figure 6.19) represent local and global (background of scatterplot) correlation, local and global (background of histogram) skewness and distribution	167
7.2	The effect of transformation for the energy variables, shown in histograms, distribution maps, local skewness maps, scatterplots (with global and local encoding) and local correlation maps	176

List of Figures

1.1	The structure of the thesis with an overview of the results, literature, contributions and publications of each stage.	7
2.1	Number of citations referring to dwelling and occupant characteristics that influence domestic electricity consumption patterns. Fig. 2 from McLoughlin et al. (2012, pp.243)	16
2.2	Spatial scale extent and resolution examples. Contains National Statistics data ©Crown copyright and database right 2015, and OS Data ©Crown Copyright and Database Right 2014	32
2.3	Simplified four stage circular process for generating a geodemographic classification. The research in this thesis focuses particularly on Stage 2. . .	38
2.4	Distribution plots for the 167 initial variables for OAC 2011 to represent variable skewness from Gale (2014b, pp.469-471). See Appendix B.8 for the variable names.	46
2.5	Compact Heatmap showing the correlation coefficient between all 41 variables of OAC 2001; using a test region of 500 OAs in four LA areas, ordered by degree of correlation.	47
2.6	Hierarchical rectangular cartogram of all OAC 2001 Super Groups in all OAs, organised by postcode hierarchy. Hue denotes OAC super-group and lightness indicates uncertainty. Fig. 5 from Slingsby et al. (2011, pp.2549) .	49
2.7	Classification maps of three urban areas: London (top), Wolverhampton (middle) and Glasgow (bottom) for one (Dataset 2) of four datasets compared for geographical distribution of clusters 6 to 8 cluster solutions for OAC 2011, Fig. 7.13 from Gale (2014b, pp.260), contains National Statistics data ©Crown Copyright and Database Right 2014, and OS Data ©Crown Copyright and Database Right 2014.	50
3.1	The responses to the prototype ‘appropriateness’ questionnaire, ranging from strong agreement (1) to the left and strong disagreement (6) to the right – Fig. 5. from Goodwin et al. (2013, pp.2521)	66
3.2	Screenshots from Smart Home HeatLines showing energy consumption for each (trial) household (vertically) over time (horizontally)	67

3.3	Screenshots from Consumption Signatures: Showing average energy consumption as Heatmap ‘Signatures’ representing 7 days (Mon-Sun: vertically) by 15 minute intervals (00:00 to 23:45: horizontally) for each appliance or group of appliances (vertically) by dataset (horizontally) . . .	69
3.4	Screenshots from Demand Horizons: Showing energy consumption as individual (small) and total (large) horizon graphs for appliances (vertically) by hour of day (horizontally)	70
3.5	Screenshots from Ownership Groups showing total and selected appliance consumption and ownership – box plots show hourly average energy consumption in grey and selected appliance(s) in black, bar chart shows percentage of appliance ownership in sample group currently ordered by type of appliance and ownership and two matrices displaying co-ownership of appliances in different ways	71
3.6	The creative design process for the smart home project. Rectangles are techniques (thick edges represent software prototypes). Concepts are round edged. Arrows show direct links between concepts and prototypes. Other links are implicit and less direct. Yellow indicates deliberate creativity mechanisms. Orange highlights processes and concepts in which creativity amongst analysts was strong. Fig. 2 from Goodwin et al.(2013, pp.2619) . .	77
4.1	The 15 Groups of Experian’s MOSAIC Public Sector Classification 2010 showing geographical locations and group descriptions. Contains National Statistics data ©Crown copyright and database right 2012, Ordnance Survey data ©Crown copyright and database right 2012. MOSAIC available via academic license from www.mimas.ac.uk , 2012	81
4.2	Three hierarchical representations of electricity consumption (colour): a. (top left) Government Regions in England containing local Authority areas – ordered spatially, b. (top right) 15 MOSAIC Groups containing 64 MOSAIC Types – ordered alphabetically, and c. (bottom) 15 MOSAIC Groups containing English Government Regions – ordered spatially	83
4.3	Hierarchical representation of electricity consumption (colour) by the 7 OAC 2001 Super Groups, containing all Local Authority regions, sized by number of electricity meters and ordered spatially	84
4.4	Design sketch for a four step application for generating geodemographics with the aid of visualisation at each stage	92
4.5	Initial Design Sketch for Stage 2: Variable Selection	93

5.1	Four methods for local statistics calculation: Each example shows Local Boroughs of Greater London, the starting location of City of London (in red) and the locality region (dark grey). Locality changes depending on the chosen method and parameters (D, N or partitioning unit). Contains National Statistics data ©Crown copyright and database right 2015, OS data ©Crown copyright and database right 2015	102
5.2	Local correlation coefficient of ‘gas consumption’ and ‘electricity consumption’ for 326 LAD in England using an adaptive moving window approach varying neighbours (N) from 100 to 50 to 25. Energy variables from (DECC, 2013a) Contains National Statistics data ©Crown copyright and database right 2015, and OS data ©Crown copyright and database right 2015	105
5.3	Four stages of the process: Input, Standardise, Locality and Output, each with two dimensions of Scale: Resolution and Extent	107
5.4	Four stages of the process for three differing geodemographics referenced in the literature	109
5.5	The data scale resolution and locality options used in the four stages of data preparation. The darker blue options were not included in the final prototype.	115
5.6	A flow diagram illustrating the data preparation process including the decisions for scale, locality and transformation	117
5.7	The four types of scale can be seen (with a black outline) in the flow of the geodemographic process prior to clustering, i.e. from Stage 1 through to Stage 3. Three types of sensitivity – scale, geography and transformation – are highlighted.	120
6.1	The Prototype Layout Panels: P1: Overview, P2: Comparison, P3: Detail. The metadata for the selected variables is shown in InfoBox in the lower third of P3	127
6.2	LAD maps from the prototype of three variables with differing values of Global Moran’s I	128
6.3	Reordering the GlobalMany view in P2 using the four reordering options in P1: Theme, Distribution, Correlation and Geography	131
6.4	The visual representation of P2 showing colour encoded global correlation coefficients values from GlobalMulti to GlobalMany	132
6.5	Asymmetrical matrix of P2 representing the local spatial and statistical visual representations of V = Multi with increased spatial aggregation as V increases reducing L from Micro to Macro	133
6.6	Use of diagonal for representing geographical distribution, statistical distribution and local skewness as interchangeable graphics	133
6.7	Demonstrating the sensitivity of the locality calculation as N in the adaptive moving window approach increases from 25 to 50 to 100 neighbours	135

6.8	Reordering the MultiMacro view in P2 with the inclusion of local statistics and the asymmetrical matrix using the four reordering options in P1	136
6.9	Panel 3 showing electric and gas central heating variables transformed via the logarithmic scale, with the original and logged distributions of the variables as histograms (top) and local (adaptive 25 neighbours) skewness maps (bottom)	137
6.10	Panel 3 showing upright and turned histograms: ‘gas central heating’ (top left) and ‘electric central heating’ (bottom right). Local correlation coefficient (adaptive 25 neighbours) is shown in the map and the scatterplot . . .	138
6.11	Default view for global values with two distribution maps shown in P3, one for each variable of the comparison, together with a large scatterplot coloured by the global correlation reflecting the P2 view	139
6.12	When local detail is displayed the local correlation map is shown in P3 with the option to display the distribution map, histogram or local skewness map (shown here) for both variables along with the scatterplot	140
6.13	Variable pair which shows a distinct difference in the local statistical values for London compared to elsewhere in the country. An example of the patterns which are identifiable when locality is included in variable selection and the geographical and statistical views are shown concurrently (in P2 and P3)	141
6.14	Variable pair with no global correlation yet clear statistical and geographical differences when locality is included. An example of the patterns which are identifiable when locality is included in variable selection and the geographical and statistical views are shown concurrently (in P2 and P3) . . .	142
6.15	Alternative map views with squares of equal size or relative to the number of regions located within	143
6.16	Increased number of data items per scatterplot when alternative SR (NUTS2, LAD and LSOA) are shown. Comparing ‘gas’ (x-axis) and ‘electricity’ (y-axis) consumption. Data items are coloured by global correlation coefficient shown in brackets	144
6.17	Scale mosaic view with superimposed colour encoded global correlation coefficients (CC) for four SR from V = Bi to V = Multi. At V = Many the variance of the four CC values is encoded, the darker the more varied . . .	144
6.18	Scale mosaic view enables the sensitivity of SR to be identified for the global correlation and skewness values for variable pairs. The prototype allows the values to be investigated for each of the four SR: NUTS2, LAD, LSOA and OA	146
6.19	Legends of Correlation Coefficient, Skewness and Standardised Variable Value	148

6.20	Grey-white-grey diverging scheme is used in P2 to avoid confusion of two diverging schemes in the histogram and central diagonal cell (sign inferred from P1) when correlation is being investigated and as the scatterplot or line glyphs when distribution is investigated	149
6.21	Transitioning from MicroBi to MacroBi using the local correlation map in P3 as an example of very small map squares and large ones causing the raster map to become more abstract	154
7.1	Histograms and maps showing standardised (left) and logged (right) statistical and geographical views for a. LAD and b. LSOA for the variable ‘travel to work by public transport’	160
7.2	P2 showing only strong (+/-0.65) correlation pairs, ordered by variance of correlation for NUTS2, LAD, LSOA and OA	162
7.3	P2 showing strong (+/-0.65) correlation values, for all four SR levels with smallest regions in the centre (a) and largest regions in the centre (b). The first 48 (of the 78) variables when ordered by variance of correlation at LAD level are shown	163
7.4	The prototype showing all seven energy variables with locality (adaptive moving window with 25 neighbours) information shown in the spatial and statistical view. The colour of the scatterplot background represents the global correlation coefficient. Four examples are highlighted and numbered.	164
7.5	Five categories of scale sensitivity shown using the scale mosaic design. Ordered by geographical hierarchy with OA as the central cell. Colour refers to the degree of positive (red) or negative (blue) correlation	172
7.6	Scatterplots with scale mosaic tiles for all energy variables in relation to average energy consumption in each of the four SR	174
7.7	MicroBi view in P3 of ‘white’ (new in OAC 2011) compared to the two ‘born in the EU’ variables (new in OAC 2011) both showing reduced global correlation compared to the previous ‘born outside the UK’ (removed from OAC 2011) variable and a more discriminating geographical variation in local correlation maps at the LAD level	179
8.1	Scale mosaic designs. Spatial: statistical boundaries (hierarchical); Temporal: seasons and yearly total (circular / hierarchical); Attribute: types of and total household energy consumption (nominal / hierarchical); and Locality: ordered number of neighbours N (ordinal).	187
8.2	Basic design for hierarchical scale mosaic designs within the raster map design of the prototype	188
8.3	Updated four stage process, removing the static classification and embedding the noClassification approach through dynamic visualisation methods and connected processes	190

Acknowledgements

First and foremost, I would like to thank my supervisor Jason Dykes for his guidance, support and critical reviews of my research throughout the process. Jason has been an inspiration to the research and I am very grateful for his commitment to not only being an excellent supervisor but a great academic mentor. I would also like to thank my secondary supervisor Aidan Slingsby, for his real interest in the topic, his positive comments, thoughtful ideas and his experiment with correlating geodemographic variables in a matrix of mini-maps.

I acknowledge City University for funding this PhD research through a Vice Chancellor's Scholarship, without which I would not have been able to pursue the research ideas or dedicate my time. I acknowledge E.ON AG International Research Initiative 2012 for funding City University and the IMDEA Energy Institute, Madrid to pursue the research project 'Visualizing the smart home: creative engagement with customer data' and for the project team for encouraging my contribution.

I would like to thank Alison Duffy, Graham Dove, Sara Jones and Amanda Brown who all made great efforts to ensure that the requirements workshop was a success. All four were present and helped to formalise the methods at the internal pilot sessions. Alison was an inspiring facilitator who ran the session professionally. Graham supported our work with the company analysts as well as led the stream with the customers. His door was always open for questions and he was been a great peer to me during the whole PhD process. I also acknowledge the work of Milan Prodanovic and Jorn Gruber from IMDEA and Veselin Rakocevic and Soroush Jahromizadeh from the Electrical and Electronic Engineering Department, who invested time and effort in building, testing and improving the smart home data model and optimisation algorithms. Particular thanks to Soroush and Jorn who devoted time to documenting and explaining the detail of their models and were both available for interviews to allow the smart home prototypes to be evaluated from a modellers' perspective. I acknowledge the visualisation design and development efforts of Aidan Slingsby, Alex Kachkaev, Iain Dillingham and Jo Wood – all of whom were encouraging of the creative and agile methods and made excellent contributions to the research project. Finally, I would especially like to thank the members of the Forward Thinking Technologies Team at E.ON and their co-workers for their interest in the project, their honest and open contributions to the requirements workshop and their positive and encouraging feedback. Their passion inspired great ideas and solutions without which the project would not have been half as successful.

The whole smart home project team were consulted on aspects of the InfoVis 2012 paper and were all helpful to its success. In particular I would like to thank Jason, Sara, Alison and Jo for their contributions to the paper and Nabiha Ahmed for her time and effort creating the video. I would also like to thank the InfoVis reviewers for their positive and critical comments which improved the paper for publication. I am grateful for Jason's time, effort and support on this paper along with other conference submissions and his help and guidance on the presentations.

I acknowledge the role of student volunteer at IEEE VIS which provided me with the financial support to attend the conference in 2012, 2013 and 2014 and allowed me to meet some really inspiring people from around the world. I also thank the giCentre and School of Informatics for further financial support to attend academic conferences and training. I thank Chris Brunsdon, Martin Charlton and Paul Harris for running the two day training session in April 2014 on 'Modelling Spatial Heterogeneity', which sparked a major change of thinking for my research. I also acknowledge Michael Sedlmair from the University of Vienna for dedicating time to discuss vPSA in the context of my research and Chris Gale from UCL for his feedback on the prototype as well as his inspiring PhD research on OAC 2011.

The encouragement of University peers and colleagues have kept me focused throughout the PhD process. I would like to thank all members of the giCentre for the useful and inspiring research discussions. In particular Aidan, Roger, Cagatay, Ali, Iain and other PhD students from A304 (including Debbie, David and Cher) for the essential breaks from research and regular lunchtime chats. I would also like to thank Jochen Schiewe, Beate Weninger, Chrisoph Kinkeldey and Jörg Münchow of the g2Lab at HafenCity University in Hamburg for welcoming me as a guest researcher, encouraging my research and helping me to stay focused and positive throughout my final year.

I acknowledge and am thankful to Tamara Munzner from the University of British Columbia and Cagatay Turkay from City University for examining my PhD thesis in February 2015 and providing valuable feedback that has improved the structure substantially. Thanks to Jason, Aidan and Cagatay for co-authoring an InfoVis 2015 paper based on the final results and to four anonymous InfoVis reviewers, whose comments improved the paper and some sections of the thesis. Many thanks also to Tim Dwyer and my new colleagues at Monash University for allowing me to spend time on improving these manuscripts.

Finally, I would like to thank all my family and friends for their support, encouragement and understanding. A huge thank you goes to my husband John for being there for me over the last few years and for completing the gruelling task of proof-reading this document. Thanks also to Tracie, Steve, Masroor and my Dad for helping John to proof-read the final version.

Publications and Presentations

The research for this thesis has been presented to academic audiences at a number of national and international conferences. The published article, short papers and posters are available in Appendix A and on-line at <http://openaccess.city.ac.uk/>.

Article

Goodwin, S., Dykes, J., Slingsby, A. and Turkay, C. (2016), Visualizing Multiple Variables Across Scale and Geography, *IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Information Visualization 2015)*, 23(01), January 2016 (*Pre-publication)

(A case study for the topology-preserving map deformation technique discussed in) Bouts, Q. W., Dwyer, T., Dykes, J., Speckmann, B., Goodwin, S., Riche, N. H., Carpendale, S. and Liebman, A. Visual Encoding of Dissimilarity Data via Topology-Preserving Map Deformation. Submitted to *IEEE Transactions on Visualization and Computer Graphics* (*Pre-publication)

Goodwin, S., Dykes, J., Jones, S., Dillingham, I., Dove, G., Duffy, A., Kachkaev, A., Slingsby, A. and Wood, J. (2013a). Creative User-Centered Visualisation Design for Energy Analysts and Modelers. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), pp. 2516-2525.

Short Papers and Posters

Goodwin, S., Dykes, J. and Slingsby, A. (2014b) Visualizing the Effects of Scale and Geography in Multivariate Comparison. Poster presented at the *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 9-14 Nov 2014, Paris, France.

Goodwin, S., and Dykes, J. (2012b) Visualizing Variations in Household Energy Consumption. Poster presented at the *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 13-18 Oct 2012, Seattle, Washington, USA.

Goodwin, S., and Dykes, J. (2012a) Geovisualisation of Household Energy Consumption Characteristics. Poster presented at the *GIS Research UK 20th Annual Conference (GISRUK)*, 11-13 Apr 2012, Lancaster University, Lancaster, UK.

Conference Presentations

Goodwin, S., Dykes, J., Slingsby, A. and Turkay, C. (2015a), Visualizing Multiple Variables Across Scale and Geography, Paper accepted to be presented at *IEEE Infovis 2015*, Oct 2015, Chicago, USA.

Goodwin, S., Dykes, J. and Slingsby, A. (2014a), Visualising the Effects of Scale and Geography in Multivariate Comparison, Presented at *Workshop on GeoVisual Analytics: Interactivity, Dynamics, and Scale* at the *Eighth International Conference on Geographic*

Information Science (GIScience 2014), 23 Sept 2014, Vienna, Austria.

Goodwin, S., Dykes, J. Jones, S., Dillingham, I., Dove, G., Duffy, A., Kachkaev, A., Slingsby, A. and Wood, J. (2013a). Creative User-Centered Visualisation Design for Energy Analysts and Modelers. Paper presented at *IEEE Infovis 2013*, 3-18 Oct 2013, Atlanta, Georgia, USA.

Goodwin, S., Slingsby, A. and Dykes, J. (2013b) Visualizing Domestic Energy Consumption of the UK, Presented at the *Annual Meeting for the North American Cartographic Information Society (NACIS)*, 9-12 Oct 2013, Greenville, South Carolina, USA.

Goodwin, S. (2012c) Geovisualisation of Household Energy Consumption Characteristics, Presented at *PhD Symposium on Household Energy Consumption*, 6 Jun 2012, University of Birmingham, Birmingham, UK.

Invited Talks

Visualizing Energy Consumption Across Scale and Space, FH-Potsdam: University of Applied Sciences, Potsdam, Germany, 1 Oct 2014.

Visualization of Geographical Data, BigData & NoSQL Meetup, Hamburg, Germany, 19 May 2014.

Visualising Energy Consumer Characteristics, g2Lab, Hafencity University, Hamburg, Germany, 15 Apr 2014.

Creative User-Centered Visualization Design for Energy Analysts and Modelers. Presented in London, UK on three separate occasions:

1. Part of 'Graphical Innovation', Unrulyversity, 27 Nov 2013.
2. Interaction Design Centre, Middlesex University, 19 Nov 2013.
3. Part of the Hochschule Karlsruhe visit, City University London, 14 Nov 2013.

Visualising Energy Consumption Characteristics, Geomob Meetup, London, UK, 29 Oct 2013.

Geovisualisation of Household Energy Consumption Characteristics, Environmental Change Institute, Oxford University, Oxford, UK, 26 Sept 2012.

Declaration

I grant powers of discretion to the university Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

Abstract

The visualisation of large volumes of data can provide rich and meaningful representations that enable users to gain insights quickly and efficiently. Household energy consumer characteristics are explored in this thesis using innovative interactive visualisation techniques. Initial research with energy analysts, from a major UK utility company, investigates visual possibilities and opportunities for future (smart home) energy analytics and explicitly uses creativity techniques for information visualisation requirements gathering. The results, along with exploratory visual analysis combining geodemographic groups and energy consumption, identifies a need for profiling consumers by typical traits. While energy consumption has been a popular topic of research in recent years, there is still limited understanding of the relationship between energy consumption and measurable characteristics of the general population. An investigation of the process of creating an energy-based geodemographic classification led to the proposal and design of a new theoretical framework for visually comparing multivariate data across scale and geography; a necessary step when selecting reliable variables for running clustering algorithms, such as during the geodemographic classification creation process.

The framework for including geography and scale in multivariate comparison forms the major contribution of this thesis. This framework is demonstrated and justified through the building of an interactive visualisation prototype, using input variables deemed relevant for consideration for energy-based geodemographic classification. Important transitions in the framework are highlighted in the proposed design, which uses both statistical and spatial representations. The utility of the framework is validated in the context of energy-based geodemographic variable selection where the multivariate geography of the UK is explored. The sensitivities of varying scale and geography – through varying resolution, extent and the calculation of locally weighted summary statistics – are investigated in context and are shown to be important elements to consider during the variable selection process. The broader applicability of the framework is demonstrated through two further scenarios where multivariate visualisation across scale and geography is shown to be important. The research provides a framework and viable solutions through which geographical visual parameter space analysis (gvPSA) can be undertaken. It uses a design science approach that results in a series of artifacts that open up new visualisation possibilities. This project covers a wide topic where the breadth of research options is extensive and many possibilities for continued research are identified.

Abbreviations

- ACORN: UK Commercial Geodemographic Product by CACI Ltd
- CPS: Creative Problem Solving
- DECC: Department of Energy and Climate Change – UK Government Ministerial Department
- DSR: Design Science Research – Academic Field
- GIS: Geographical Information Systems/Science – Computer System for Geographical Data/Academic Field
- GISRUK: GIS Research UK – Annual Conference Series
- IEEE: Institute of Electrical and Electronics Engineers – Professional Association
- Infovis: Information Visualisation – Academic Field and Conference Series IEEE InfoVis
- LAD: Local Authority District – UK Administrative Geographical Regions
- Log: Logarithmic Scale – Mathematics
- LSOA: Lower Level Super Output Areas – UK Census 2001 & 2011 2nd Tier Statistical Regions
- MAUP: Modifiable Areal Unit Problem – source of statistical bias when aggregating data
- MOSAIC: UK Commercial Geodemographic Product by Experian Ltd
- MoSCoW: Must, Should, Could, Won't – Requirements Prioritisation Method
- MSOA: Middle Level Super Output Areas – UK Census 2001 & 2011 3rd Tier Statistical Regions
- NACIS: North American Cartographic Information Society inc. annual conference series
- NUTS1: Nomenclature of Territorial Units for Statistics – Tier 1 European Statistical Regions
- NUTS2: Nomenclature of Territorial Units for Statistics – Tier 2 European Statistical Regions
- OA: Output Areas – UK Census 2001 & 2011 1st Tier Statistical Regions
- OAC: Output Area Classification (UK) 2001 and 2011 by ONS (distinguishable by date)

- ONS: Office of National Statistics – UK Government Department
- PCA: Principal Components Analysis – Statistics
- SciVis: Scientific Visualisation – Academic Field
- SPLOM: Scatterplot Matrix – Visualisation Method
- TVCG: (IEEE) Transactions on Visualization and Computer Graphics – Academic Journal
- Vis: Visualisation – Computer Science Discipline – and Annual Conference Series IEEE VIS
- vPSA: Visual Parameter Space Analysis
- UCL: University College London

Thesis Specific (See sections in brackets for further definitions)

- D: Distance Measurement in Fixed Moving Window Locality Calculation (5.1)
- gvPSA: Geographical Visual Parameter Space Analysis (2.4.5)
- IE: Input Scale of Extent (5.3.1)
- IR: Input Scale of Resolution (5.3.1)
- L: Number of Local Summary Statistics (5.1 and 6.1)
- LE: Locality Scale of Extent (5.3.1)
- LR: Locality Scale of Resolution (5.3.1)
- MQ: Motivational Question (1.2)
- N: Number of Neighbours in Adaptive Moving Window Locality Calculation (5.1)
- OE: Output Scale of Extent (5.3.1)
- OR: Output Scale of Extent (5.3.1)
- RQ: Research Question (1.2)
- SE: Scale of Extent (5.2)
- StE: Standardised Scale of Extent (5.3.1)
- SR: Scale of Resolution (5.2)
- StR: Standardised Scale of Resolution (5.3.1)
- US#: User Story (Numbered) (4.2.2)
- V: Number of Variables (6.1)

1

Introduction

This thesis uses innovative interactive data visualisation to study large datasets that contain information relating to the behaviour and characteristics of energy consumers. It does so by applying creative design techniques within the constraints of established approaches to visual design. This results in a series of artifacts that open up possibilities for using visualisation in new ways with the particular focus on facilitating geodemographic variable selection, where multiple variables are used to represent the variation of the population and its behaviour across geography.

To introduce the topic, this chapter outlines: the research context, the research aims, objectives and motivation, the research questions, scope and ethics, along with the methodologies used. A textual and visual overview of the thesis structure is presented, relating each phase of research to the relevant chapters and research questions.

1.1 Research Context

Growing environmental pressure to meet ambitious targets to reduce worldwide carbon dioxide (CO²) emissions increases the need to address current consumption levels and to better understand consumer habits and behaviour. Despite extensive academic research in the area, there is still limited understanding of the relationship between household energy consumption and measurable characteristics of the general population (Druckman and Jackson, 2008). The research in this thesis investigates possibilities and

opportunities to improve the current data analysis and the visualisation of energy consumer characteristics. Data visualisation and visual analytics offer real opportunities within the energy domain; from network analysis and grid analytics through to the representation and analysis of energy consumption patterns for both the householder and the energy supplier. Opportunities are rapidly growing in the industry given the vast quantities of data which are set to become available from smart technologies (Rusitschka et al., 2010). These new datasets allow for more advanced analysis to improve understanding of consumption patterns and consumer behaviour (Firth et al., 2008) and optimise the management of supply and demand (Clastres, 2011). Due to these changing technologies and emerging (smart) data sources, it is also evident that there is a need to allow for adaptable and flexible analytical processes.

The context of energy consumer characteristics is one example of multivariate geographical data analysis where the research in this thesis is applicable. Motivation for investigating the energy industry in particular came from the author’s previous work experience, analysing energy tariff and consumer characteristics¹. This background knowledge in the industry prompted open questions which motivated the research, such as “*What is the future for household energy analysis?*” and “*What value can be derived from energy consumption data through data analysis and visualisation?*” These broad questions were addressed at a requirements gathering workshop with energy analysts from a major UK energy provider during the smart home research project (outlined in Chapter 3). The workshop included the explicit use of *Creativity Techniques* to stimulate open discussions and creative thinking. Many aspirations and ideas for the future of energy analysis were identified at this workshop. One aspiration in particular – “*I would like to know the who, what, when, where and why of energy consumption*” – motivated much of the research in this thesis. The *what?* and *when?* are visually investigated for the smart home project, where four visualisation prototypes, revealing patterns of household appliance use over time, are evaluated with the energy analysts participating in the initial workshop (as discussed in Section 3.5). The *who?* and *where?* of consumption forms the focus for Chapter 4 and builds the case study for the remaining research, where geodemographics in the context of energy consumer profiling are investigated.

An investigation and analysis of the variables within UK geodemographic classifications was included in the author’s previous research (Goodwin and Dykes, 2008) and this knowledge prompted the third motivational question for this research: “*Is there a need for an energy-based geodemographic classification?*” Relevant research relating to domain-specific geodemographic classification is discussed in Section 2.3 where the

¹The author worked as an Business Intelligence analyst for three years, from 2008-2011, for Verivox GmbH – the largest energy price comparison company in Germany

process for generating domain-specific geodemographics is described (in Section 2.3.3) and simplified into a four stage process for this research (in Section 2.3.4). This discussion continues in Chapter 4 with specific focus on Stage 2 of the four stage process – the variable selection stage. During this stage decisions are made based on variable structure, where heavily skewed or strongly correlated variables can bias the clustering process. Data scale, geographical variation and data transformation are all shown to be important when deciding which variables to select for clustering. The complexities and sensitivities associated with the representation of geographical variation of data variables statistically through the use of local (geographically weighted) statistics as well as those associated with varying the data scale are described and demonstrated (Chapter 5). A framework to visualise the complex parameter space associated with comparing multivariate data across geography and scale forms a major academic contribution of this thesis (described in Chapter 6). Visual representation possibilities are identified for different types of data (Section 6.2). An interactive visualisation prototype is designed, developed and described to demonstrate the feasibility and effectiveness of the framework (Section 6.3). This stage of the research can be linked to visual parameter space analysis (vPSA), recently described as interactive visualisation that facilitates parameter space analysis (PSA) (Sedlmair et al., 2014). Whilst vPSA explores the effects of varying parameter values in a model’s parameter space, this research explores the effects of varying the parameters of geography and scale for geodemographic variable selection. Techniques and terminology from vPSA, such as parameter sensitivity analysis (Sedlmair et al., 2014), are used in the context of this research, particularly in Chapters 7-9. The geographical visualisation techniques necessary to present the parameters expands vPSA to geo-visual PSA (shortened to gvPSA).

The prototype is designed within the context of energy-based geodemographic variable selection, where the effects of varying scale and geography are investigated. The prototype is validated through an investigation of this scenario (outlined in Chapter 7), where visual analysis highlights the sensitivity of scale, geography and transformation to the variable selection process. The research demonstrates that well-designed interactive visualisation can provide access to the most important characteristics of the parameter space and in particular provide information about its geographic variation. Further applicability of the framework is demonstrated through additional designs (Section 8.1) and the preliminary investigation of two additional scenarios (Section 8.3). The research with energy analysts (described in Chapter 3) demonstrates the opportunity for real time engagement with emerging smart home variables through interactive visualisation. This provides the second scenario ‘Smart Home Analytics’ in which the framework is shown to be applicable (Section 8.3.1). The third non-energy scenario ‘Survey Response

Modelling’ demonstrates the value of the framework for a current research project where survey nonresponse bias is investigated through auxiliary data sources (Section 8.3.2). All three scenarios demonstrate that the framework and prototype designs are useful and applicable to real-world problems.

Although visualisation is shown to be beneficial to variable selection and the generation of geodemographics, this research also reveals another approach where static classification is replaced by interactive visualisation as an alternative means of understanding dynamic multivariate geographic data in applied contexts. This dynamic approach is termed *noClassification* in this thesis, as geodemographic classification does not actually occur but the selected (up-to-date and potentially near real-time) variables are grouped, analysed and interpreted for decision-making through interactive and dynamic geo-visual analytic approaches. This *noClassification* approach is discussed further in Chapters 7-9.

The research in this thesis is presented in the context of design science research (DSR – discussed in Section 2.2.2), where the prototype is an instantiation of the model represented by the framework and it is used to explore the utility and validity of the framework in an applied context (Hevner and Ram, 2004). The exploration of Scenario 1 (Chapter 7) demonstrates that visualisation is beneficial to the variable selection process. The results confirm the need for a more visual and dynamic approach to the process of generating geodemographic classifications (as suggested in Section 4.2), as opposed to the more static approach in which classifications are currently built. This thesis complements a trend in geodemographics towards more local and domain-specific classifications (Singleton and Longley, 2009b). The research also demonstrates that there are many complexities and uncertainties involved. It is known that well-designed interactive visualisation can present multivariate data in a richer, more nuanced way than is achieved through static classifications (Slingsby et al., 2011). The visual approach challenges current practice and is both supportive of insight generation and well suited to the dynamic way in which data is generated.

In accordance with many research projects, a broad initial goal is refined to a very specific problem area. As the research for this thesis developed and became more focused, the original goal of producing an energy-based classification and associated visual representation was adapted to a research goal which concentrated on visualising the complex parameter space of multivariate data across geography and scale. The newly focused goal introduced two additional research questions (RQ3 and RQ4). RQ2 was also modified from “*How can data visualisation aid the process of generating a geodemographic classification*” to “*How can data visualisation aid the variable selection process?*”(see Section 1.2). Wood et al. (2014) describe this as a new ‘channel’ of

research and their paper confirms that in visualisation research a new direction or strand of research can often develop when working together for a long time with one client or continued research into one specific domain.

In terms of academic audience, the research in the thesis is relevant to energy analysts investigating new methods for the visualisation of energy consumer characteristics, geographers or statisticians researching domain-specific or geographically weighted geodemographics and classification creators or data engineers interested in visual methods for multivariate comparison with the inclusion of data scale and geography.

1.2 Aims and Research Questions

This thesis aims to investigate innovative interactive data visualisation methods to allow for better exploration of the characteristics of energy consumers through large datasets. Three open and broad questions, prompted from previous academic and professional experience, motivated the research:

- *MQ1: What is the future for household energy analysis?*
- *MQ2: What value can be derived from energy consumption data through data analysis and visualisation?*
- *MQ3: Is there a need for an energy-based geodemographic classification?*

These are investigated through an academic literature review in Chapter 2, and continued through reflecting on the requirements of energy analysts in Chapter 3 as well as the exploratory visual analysis of energy use and an investigation of the geodemographic process in Section 2.3 and Chapter 4. The research specifically focuses on four research questions (RQ):

- *RQ1: Which demographic or socio-economic variables should be combined with energy consumption variables to enable characteristics of UK household energy market to be identified?*
- *RQ2: How can data visualisation aid the variable selection process?*
- *RQ3: What are the sensitivities and uncertainties associated with variable scale (spatial, temporal and attribute), geography or transformation in multivariate comparison?*
- *RQ4: For spatial scale (resolution and extent) in particular: how can data visualisation expose these sensitivities and uncertainties when comparing multivariate areal datasets?*

1.2. AIMS AND RESEARCH QUESTIONS

Relevant research relating to each RQ is investigated in Chapter 2. Exploration of RQ1 is continued in Section 4.3 with research into possible candidate variables for energy classification and the selection of variables for further analysis. The investigation of RQ1 continues in Chapter 7 with the demonstration and validation of the prototype. RQ2 is addressed in Chapters 4-7 with the investigation of variable relationships and visualisation possibilities. RQ3 is addressed in particular in Chapter 5 and continued with the investigation of RQ4 in Chapters 7 and 8

In summary, RQ1-4 are discussed in Chapter 2, outlined in the context of variable selection for geodemographics in Chapters 4-8 and combined and discussed in detail in Chapter 9 in combination with the motivational questions and reflection of the needs of the energy analysts' from Chapter 3. Figure 1.1 visually represents the structure of the thesis in detail, illustrating how each RQ links to each stage of research.

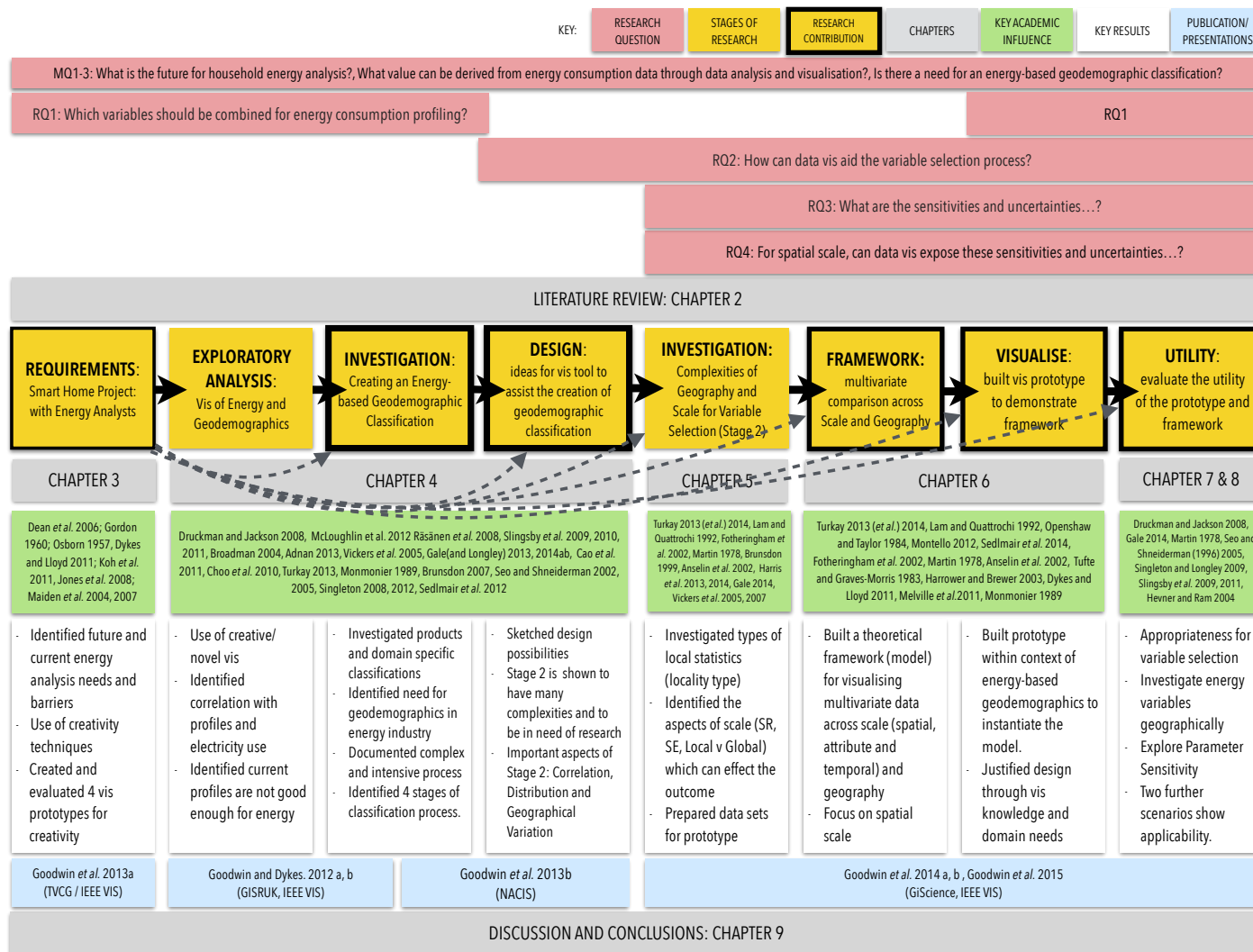


Figure 1.1: The structure of the thesis with an overview of the results, literature, contributions and publications of each stage.

1.3 Thesis Scope and Contributions

While the initial aims of the research project are relatively broad, its scope focuses on the creation of a theoretical framework for comparing multivariate data, with the consideration of scale and geography. An interactive prototype with clear justification of design decisions is built to visualise the framework. The framework is validated through the in-depth investigation of the applicability to variable selection for energy-based geodemographics (Chapter 7). Further applicability is demonstrated through the preliminary investigation of two additional scenarios (Section 8.3). The research opens up a new design space and paves the way for continued research in the area.

The research outlined in this thesis makes a number of contributions to a wide research community:

- The new theoretical framework for visualising both scale and geography in multivariate comparison explained in Section 6.1 is the primary contribution of this thesis;
- Visual designs used in the prototype to demonstrate multivariate geographical comparison – in particular, the scale mosaic design – explained in Section 6.3 and Section 8.1 form a further primary contribution;
- Results of exploration and sensitivity analysis of variables relating to household energy usage in Chapter 7 are research contributions that relate in particular to the UK energy industry;
- Research and designs to improve transparency of the process of generating a geodemographic classification through visual and iterative applications throughout the thesis (in particular Chapter 4 and 7) are contributions to geodemographic research;
- Considering geography as an input to visual parameter space analysis to establish geo-visual PSA forms a further contribution to research in the area of vPSA;
- The use and evaluation of creativity techniques for information visualisation requirements gathering in Chapter 3 is an additional contribution, which opens up opportunities for using creativity techniques to develop visualisation requirements, applications and design;
- Finally, the link between visualisation research to design science research terminology adds a further minor contribution of this research.

These contributions to visualisation research are described as design science artifacts, known as constructs, models, methods and instantiations. Using known *constructs* of scale (SR and SE), geographically weighted statistics and vPSA, a new *model* is built through the framework and the prototype is an *instantiation* of this model. The research demonstrates some new *methods* related to visual design (scale mosaics) and analysis techniques (creativity). The framework together with its instantiation demonstrate the potential for the new *constructs* of *gvPSA* and *noClassification* in this context and demonstrates applicability in other scenarios and future work.

1.4 Research Methods

The research project combines many research methods including creation and research methodologies (Oates, 2005), exploratory data analysis (Tukey, 1977), design science (Hevner and Ram, 2004), visualisation design, qualitative and quantitative analysis methods, including geographically weighted statistics (Fotheringham et al., 2002) and spatial auto-correlation (Anselin et al., 2002) and the use of creativity techniques (Dean et al., 2006), rapid prototyping (Dow et al., 2010), prioritisation methods and agile development methodologies (Cohn, 2005). The research also builds on the visualisation knowledge base by designing a new theoretical framework for the visual comparison of multivariate data across scale and geography. Terminology from design science research (Hevner and Ram, 2004) is used to justify the approach, identify artifacts and demonstrate research contributions.

1.5 Use of Academic Literature

In order to identify relevant literature, a semi-systematic literature strategy was used for the initial search. Keywords were searched in digital libraries such as ScienceDirect², IEEE Xplore³, ACM Digital Library⁴ and GoogleScholar⁵. Decisions regarding the relevance of the literature were made based on the title and abstract, the date, journal name, knowledge of the author(s) and by browsing the citation list for relevant references. Relevant papers were read and notes were taken. All citations were stored and backed up using the stand-alone version of Zotero⁶, where collections of literature were distinguished and notes were saved with the citation for future reference. Additional literature was found through relevant citations and subsequently assessed for relevance. Some fruitful keyword searches were saved and email notification activated (e.g. with GoogleScholar) when new citations became available. Further research was discovered through collaboration with the Centre

²<http://www.sciencedirect.com/>

³<http://ieeexplore.ieee.org/>

⁴<http://dl.acm.org/>

⁵<http://scholar.google.com>

⁶www.zotero.org

for Creativity in Professional Practice at City University London, where relevant research was highlighted. In addition, more recent research was identified at conferences or from reading conference proceedings, in particular IEEE VIS, GISRUK and GIScience from 2011-2014.

1.6 Research Ethics

The research outlined in Chapter 3 reports on part of a research project undertaken by City University London and the IMDEA Energy Institute, Madrid and was funded by E.ON AG International Research Initiative (IRI) 2012. The research was a collaborative project involving a number of academics from the IMDEA Energy Institute and the departments of Engineering, Human Computer Interaction and Informatics, in particular members of the Centre for Creativity in Professional Practice and the giCentre at City University London. There were four separate phases of the research project, of which the second stage (which involved creative visualisation for the energy company) was undertaken by the giCentre and was led by the author. The requirements, techniques and results relevant to this thesis are outlined in Chapter 3.

This research involved three external workshops with energy analysts from the Future Technologies Team at E.ON UK. Prior to each of the workshops, the research project and the PhD research was explained to the participants and a specifically designed consent form was signed by each participant (as shown in Appendix B.1). For the results to be used in this thesis and other academic publications (e.g. Goodwin et al., 2013) all the discussions, quotes and ideas recorded (both written and audio) or photographed during the workshops were anonymised and no names are mentioned or published.

This research involved the use of two datasets. The first was a modelled dataset created specifically for the E.ON research (Gruber and Prodanovic, 2012), which does not involve any individual data, but instead was built using an openly available survey of appliance use (Zimmermann et al., 2012). The second data source was from a smart home trial project by E.ON. Permission and rights to use this data for academic purposes formed part of the project contract. The data structure is only described in general terms in this thesis, in order to ensure that the members of this trial are not identifiable and that there is no breach of contract. Visual representations are of aggregated data where no characteristics of the individual households are identifiable except very broad trends and no direct insights relating to the trial data or households are discussed or described.

The rest of the data used in this thesis uses publicly available areal (aggregated) datasets of demographic, housing and socio-economic based variables. Although ethical considerations need to be taken into account, the use of aggregated data removes the ability to identify individuals and means that ethical issues, such as those relating to disclosure,

are minimised. The ONS Census variables and the DECC energy-based variables used for the research for this thesis (in particular in Section 4.1 and Section 5.4) are aggregated to areal units and disclosure issues are dealt with prior to the release of the data; small numbers are automatically removed prior to publication (DECC, 2013a).

In summary, to reduce the risk of disclosure, all data relating to households was aggregated during or prior to the research and the use of data anonymisation is used to protect the individuals involved in the workshops.

1.7 Thesis Structure

The structure of the document is illustrated in Fig. 1.1. The thesis has nine chapters:

Chapter 1 introduces the topic, background, scope, contributions and limitations of the research.

Chapter 2 investigates, describes and evaluates the academic literature relating to energy data analysis, data visualisation, geodemographic profiling and multivariate comparison. Four stages for generating a domain-specific geodemographic classification are introduced, with particular focus on Stage 2 – the geodemographic variable selection process. The relevance and importance of creativity and design science research are also introduced and discussed within the context of data visualisation.

Chapter 3 describes a collaborative research project, which investigates smart home data analysis and visualisation with energy analysts. Creativity techniques are used within the requirements gathering workshop, where requirements for the future of energy consumption analysis and visualisation are explored. Four visualisation prototypes, linked to the *what* (appliances) and *when* (time of day) of energy consumption, were designed and their appropriateness evaluated in combination with the degree of creativity within the design process.

Chapter 4 continues the investigation by focusing on the *who* and *where* of energy consumption. Exploratory visual analysis of geodemographics with energy consumption is described, tools for creating domain-specific geodemographic classifications are explored, a visual approach to improve the generation process is proposed, user stories are identified and preliminary visual design ideas illustrated. Complex multivariate and multidimensional comparison is identified as key to the variable selection stage of the process. Potential candidate variables for classification are explored, described and defined for continued investigation.

Chapter 5 identifies the dimensions and complexities of varying data scale (resolution and extent) and the calculation of local geographical variation within the context of variable selection for geodemographics. Candidate variables are prepared for multivariate comparison. Methodologies are justified through the knowledge of the

chosen variables, the context of energy-based geodemographics and the statistical and geodemographic literature.

Chapter 6 describes and illustrates a framework for the visual representation of the complex parameter space of comparing multivariate data across scale and geography. Design decisions for an interactive prototype built to demonstrate the framework – with the specific focus on spatial scale resolution within the context of energy-based geodemographics – are described and justified with reference to the visualisation domain and relevant examples.

Chapter 7 demonstrates the value of the framework and prototype through the exploration of variable selection for geodemographics (scenario 1). The prototype design is shown to be appropriate to the context through linking to user stories (from Section 4.2.2). The utility is demonstrated through the exploratory visual analysis of the energy variables and parameter sensitivity of scale, geography and transformation.

Chapter 8 describes the broader application of the prototype and framework through the demonstration of possible designs for future work and the explanation of two additional scenarios Smart Home Analytics and Survey Response Modelling.

Chapter 9 combines the research outlined in Chapters 3-8 and discusses the research with reference to the research questions outlined in Chapter 1 and literature described in Chapter 2. The scope of the research is described, the research findings summarised and concluded and research contributions outlined.

Additional appendices are attached. **Appendix A** includes published papers and posters and **Appendix B** includes additional clarification and justification of the work.

2

Literature Review

This thesis investigates the use of data visualisation for analysing household energy consumer characteristics. This broad research area draws on academic literature from many disciplines including engineering, statistics, geography, social sciences and human-computer interaction. The aims, research questions and scope of the thesis (outlined in the previous chapter) focus the review of relevant research. This is structured in the following sections: Energy Consumption Analysis, Data Visualisation, Geodemographic Profiling and Multivariate Comparison.

2.1 Energy Consumption Analysis

This section describes recent research relating to the analysis of household energy consumption data in the UK and the current improvements to the grid network and new technologies, which are leading to a smarter, more efficient and data abundant industry. Due to these changes there are growing needs and opportunities for advanced data analysis to improve knowledge and understanding of the datasets throughout the industry, from analysing the energy network down to better informed consumers.

2.1.1 Household Consumption Analysis

The UK has a target to reduce CO² emissions by 60% compared to 1990 levels by 2050 (McLoughlin et al., 2012). To reach such a target, reduction measures must be tackled in all energy sectors from industrial through to domestic use. In 2013, total con-

sumption from the domestic sector (excluding transport) was 43.8 million tonnes of oil equivalent (mtoe) remaining stable compared to 2012 (43.7 mtoe). This measures 29% of the UK's total electricity consumption (DECC, 2014). There has been significant technical improvements in the industry over recent years, including low-energy appliances, smart metering and micro-generation (DECC, 2014). However, to reduce consumption levels to the significant degree required to meet the national targets, action must involve not only technical improvements but effective policies, regulations of utilities, the reform of tariffs and much greater information for the consumers (Boardman, 2004).

There has been extensive research from academia, government and from within the industry, to better understand the key drivers of household energy consumption in order to facilitate a change in consumer behaviour. The UK Department of Energy and Climate Change (DECC) produces an annual report providing an overview of the trends and key drivers that influence household energy consumption in the UK, investigating trends in the data from 1970 to the current year (although detailed data starts from 2008) (DECC, 2014). Space heating is the major consumer, accounting for 58% of all delivered energy consumed in 2000. Although outside temperature has a major influence on year-to-year fluctuations, this percentage has been increasing since the 1970s despite the increased presence of household insulation. This rise in consumption is due to the growth of central heating, increased population and a rise in internal temperatures (DECC, 2014). When combined with water heating, heating accounts for 82% of all household use. Other major areas of consumption are lighting, appliances and cooking. The increased number of home appliances and the change in society is reflected by a 157% increase in consumption from lighting and home appliances from 1970-2000 (DECC, 2014).

Appliance use and consumption research covering 12 European countries reveals that, by changing to the best available technology and altering behaviour, potential savings are estimated to be 1300 kWh of electricity per household per year (de Almeida et al., 2011). Cold appliances, lighting and desktop PCs (including monitors) are the main appliances responsible for the savings of 26.8%, 23.6% and 10.8% respectively. This research indicates that more regulation changes, more informed consumers, better labelling and more suitable financial incentives may be effective in stimulating market transformation (de Almeida et al., 2011). A more recent study of 251 households in England reveals that the total electricity saving per household range from 491 kWh to 677 kWh depending on the type of household. This is expressed as a minimum value, as the lighting saving was underestimated (Zimmermann et al., 2012). This value is far lower than the saving stated by de Almeida et al. (2011), indicating that, the research may have been more detailed, or that there is simply less of a saving in England in comparison to other countries in Europe. Despite the difference in value, the detailed recommendations are

similar; cold appliance replacement can save up to 358 kWh/year per household and switching a desktop PC to a laptop could save up to 128 kWh/year. Investigation of the standby option on appliances also highlights that enforcing a maximum standby power of 0.5W could reduce consumption by 111 kWh/year per household.

Despite increased research in the area and the better understanding of where savings can be made, our understanding of the relationship between household energy consumption and measurable characteristics of the general population remains limited. Research shows that there are many social and environmental factors that may influence energy use. The study by Zimmermann et al. (2012) shows that use depends on household type, with floor size, house type and number and type of resident showing differing energy usage. Relevant research by Druckman and Jackson (2008) reports that UK domestic consumption is strongly related to disposable income level with other highly influential factors being household composition, dwelling type, tenure and urban/rural location. A subsequent literature review by McLoughlin et al. (2012) confirms that in terms of citations; household (rather than disposable) income and dwelling type appear most frequently in the academic literature as having a key influence on domestic electricity consumption, with number of occupants and appliance holdings being the next most frequently cited (see Fig. 2.1). It is, however, noted that frequency of citations is also likely to be a consequence of data availability and highly relevant variables such as floor area, appliance rating and time of use may have been overlooked in some literature (they appear mid-graph in Fig. 2.1) due to difficulty in obtaining the datasets. Proxy datasets have often been used, for example ‘number of rooms’ which is surveyed in the UK Census is often used to approximate ‘floor area’, a variable which is not surveyed (Yohanis et al., 2008). This research, along with others (such as Semenik et al., 1982; Dillahunty et al., 2009; Dillahunty and Mankoff, 2011) demonstrates that domestic consumption correlates with many socio-economic and geographic characteristics of the population, and continued research in this area is necessary to increase understanding of the complex variations in domestic consumption across the country.

2.1.2 Smarter Energy Analysis

In recent years the energy industry has seen substantial technological improvements which are enabling the move away from a standard, passive and inefficient power grid network to the delivery of a *smart grid* linked by sensors, monitors and other digital technologies. This will allow for better communication and more efficient control of energy supply and demand (Giordano and Fulli, 2012; Clastres, 2011). *Smart meter* technology allows for energy consumption to be recorded at frequent intervals (e.g. hours or minutes) and data

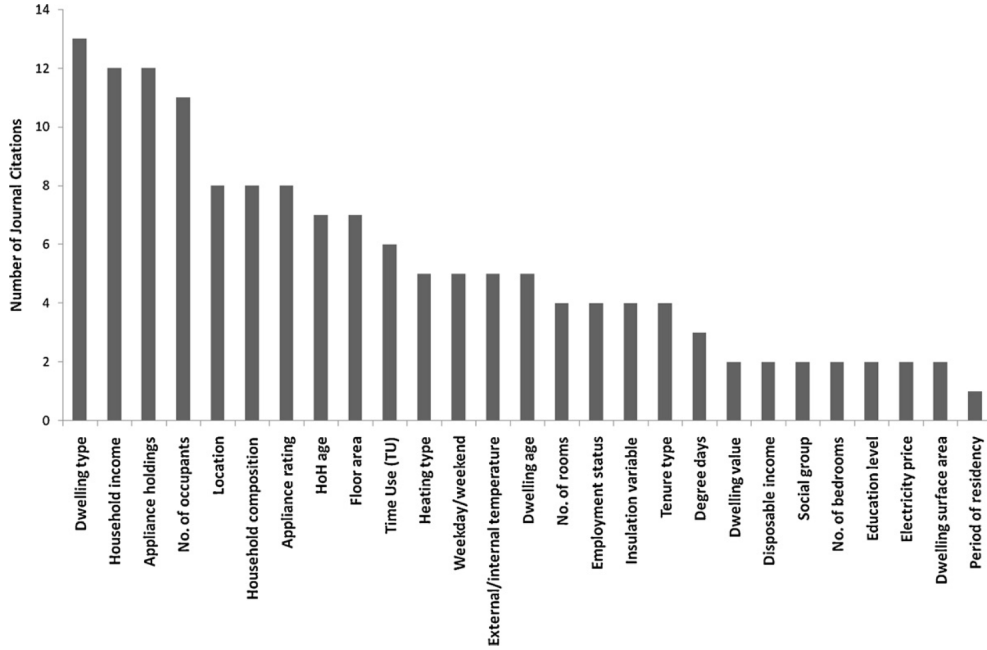


Figure 2.1: Number of citations referring to dwelling and occupant characteristics that influence domestic electricity consumption patterns. Fig. 2 from McLoughlin et al. (2012, pp.243)

to be reported back to both energy supplier and consumer allowing for near real-time feedback of energy use (Darby, 2010; Jennings, 2013).

Smart meters are expected to greatly improve user awareness and allow for the regulation of energy consumption at the household level (Darby, 2010). A study of early adopters by Hargreaves et al. (2010) reports improved awareness, but further studies are needed to identify whether changes in behaviour are long term and to understand the differences across household types. Many EU countries have started to introduce smart meter technology into households, with Italy reaching 85% household coverage in 2010 (Clastres, 2011). In 2009 the UK Government announced the intention to introduce smart meters into all households for both electricity and gas by 2020 (Faruqui et al., 2010).

In addition to new meters, some homes are being equipped with sensing systems and control devices capable of communicating with each other to enable more efficient management of energy consumption and associated costs as well as increasing the levels of security and comfort within the home (De Silva et al., 2012). Advances in smart meter and smart home technologies are becoming increasingly important to energy suppliers and consumers and the vast quantity of data yielded is increasing the value of data in the industry exponentially (Rusitschka et al., 2010). Energy data analysts and modellers are beginning to explore opportunities to use the emerging data to understand consumption patterns and consumer behaviour (Firth et al., 2008) as well as optimise the management of supply and demand by deriving flexible tariffs and services (Clastres, 2011).

The new data sources are popular in recent research, where advanced data analysis techniques are utilised to investigate load-profiling and the classification of consumer appliance use (such as Fan et al., 2012; Lühr et al., 2007; Asare-Bediako et al., 2011; Weiss et al., 2012). With regard to household characteristics, McLoughlin et al. (2012) investigate data from a smart meter trial of 4200 households in the Republic of Ireland in combination with dwelling, demographic and socio-economic characteristics of each household. This research reveals that *“Electricity consumption patterns for domestic dwellings are highly stochastic, often changing considerably between customers”* (McLoughlin et al., 2012, pp.240). The detailed analysis allows for the investigation of household characteristics on total electricity demand as well as load profile properties such as time-of-use (McLoughlin et al., 2012). In terms of total electricity consumption, the results reflect the literature mentioned in the previous section, that consumption is strongly influenced by number of rooms (used as a proxy for household size), type of dwelling, age of the head of household and social class, with ‘Higher Professionals’ consuming more than others and inferring a link to income. The investigation of energy load profiles showed high influences with household composition, number of bedrooms, water heating and cooking type and time-of-use was heavily influenced by occupant rather than dwelling characteristics (McLoughlin et al., 2012). This project is unique in recent research, which often uses much smaller household samples, as the adoption of smart meters across the population at present is limited in most countries. In order to predict what the future of smart home data can bring, data modellers are beginning to use these initial datasets and large household surveys (such as Zimmermann et al., 2012) to simulate data from thousands or millions of homes (e.g Gruber and Prodanovic, 2012).

2.1.3 Energy Consumption Analysis Summary

Over recent years, the amount of energy consumed by households has been of growing interest to individuals, organisations and governments. Reducing domestic energy consumption is particularly challenging due to large variations in household energy use. Most recently we have seen an advancement in smarter technologies with smart meters, sensors and monitors allowing data to be reported back to the supplier and consumer in near real-time. This is seeing a vast increase of data in the industry and energy analysts and modellers are beginning to investigate the possible data opportunities.

2.2 Data Visualisation: Design, Application & Creativity

This section introduces the domains of data visualisation, visual analytics, information visualisation (infovis), geographical data visualisation (geovis) as well as the importance of design decisions and creativity.

2.2.1 Visualisation

In recent years, the vast improvement and reduced cost of technical equipment, from personal computers through to digital sensing technology, has lead to a digital data boom. Many more datasets are being monitored and analysed to identify patterns, trends and anomalies. The analysis of data is *“to highlight its features in order of their importance, reveal patterns, and simultaneously show features that exist across multiple dimensions”* (Fry, 2008, pp.1). Both during and after analysis, effective graphical representation of these features has been shown to be useful and important in enabling users to gain the bigger picture quickly and efficiently.

In order to effectively represent large datasets visually, there are decisions to be made based on reducing the data through filtering (the elimination of elements), aggregation (the grouping of elements), or a combination of both (Munzner, 2014). This may involve statistical means such as clustering, sampling or data transformation, or simply the aggregation or filtering of data by attribute type, by period of time or by geographical location. The decision depends on the dimensions of the data and the tasks of the users. Representing the same dataset from several perspectives using multiple views can also be used to reveal hidden meaning and relationships in the data being analysed. Few (2009) encourages visualisation designers to explore the data through the use of faceted analytical displays (differing from dashboards which are used specifically to monitor, not analyse the data) in order to ensure that the *“meaningful relationships that exist in our data”* can be explored (Few, 2009, pp.104).

Visualisation is described as both an academic Art and a Science and it overlaps with exploratory data analysis, statistical graphics, information design, information visualisation (infovis), information graphics (infographics), cartography, scientific visualisation (scivis) and visual analytics. Of these disciplines infovis is often defined as; *“the use of computer-supported, interactive visual representations of data to amplify cognition”* (Card, et al. 1999 in Fekete et al., 2008, pp.2), while visual analytics tends to use larger datasets with the aim of making the electronic processing of data more transparent by combining the strengths of humans with machines (Keim et al., 2010). In all fields the design of the visualisation is of great importance. Ware (2013, pp.3) states that *“One of the greatest benefits of data visualization is the sheer quantity of information that can be rapidly interpreted if it is presented well”*. The active

visualisation research community is helping to answer how we define *well-presented*, with continuous advances in visualisation theory and best practice, where principles of design have proved to be of particular importance.

2.2.2 Visual Design

Visualisation design decisions involve trade-offs, which depend on the needs of the user and the task at hand. These design decisions include deciding which features to highlight or emphasise, using which *visual variable*. Bertin (1967) introduced the original visual variables for graphic design as *position*, *size*, *shape*, *value*, *colour*, *orientation* and *texture*. The foundations of visualisation design are heavily linked to graphic design research from Bertin (1967) and Tufte (1983; 1991). Subsequent research adapts these foundations for specific examples, such as the ranking of visual variables depending on accuracy for perceptual tasks for nominal, ordinal or qualitative data types (Mackinlay, 1986). This included amendments to the original visual variables, for instance dividing *colour* into *colour hue* and *colour saturation*. Visualisation designers must take into account the user tasks and data types but also the interactions available and the limitations of the display size, the user and the computational device (Munzner, 2014). The design decision stages have been researched within the visualisation community with specific models created to aid the visualisation creation and evaluation process, such as the nested model (Munzner, 2009) and the nine-stage framework (Sedlmair et al., 2012).

The visualisation of spatial data with an emphasis on the geographical elements of the data, is often termed geographical information visualisation or geovisualisation (geovis) (Dykes et al., 2005). Certain aspects of the visualisation design and interaction need to be specifically tailored to reflect the geographical element of the data and some visual variables need to be considered differently, such as the careful use of colour schemes (Harrower and Brewer, 2003) and position (Wood and Dykes, 2008). Geography is an important element of the data compared and visualised in the research for this thesis and this review will draw on more research in this area, as well as other visualisation domains, in future sections (and subsequent chapters).

The design space of visualisation possibilities is vast and the choice of which representation to use reflects the design and data reduction decisions previously mentioned. Visualisation is becoming more accessible with improved tools, increasing prevalence of software development skills and increasingly availability of open data. This has seen an increasing number of artists and designers applying typical infovis principles in “*powerful and even artistic means of expression*” (Vande Moere and Purchase, 2011, pp.356). Many visualisations are being presented in exhibitions and published as attractive coffee-table books (e.g. Bohnacker, 2012; McCandless, 2012; Cheshire and

Uberti, 2014), which could be described as beautiful, novel or creative representations of data. Measuring the success of a visualisation solution is, however, subjective and highly complex. Vande Moere and Purchase (2011) described this process as being a balance of the three requirements of attractiveness, soundness and utility. *Attractiveness* relates to the aesthetics – the appeal and beauty of the given solution, which links to the originality, innovation and novelty of the user experience. Just as important is *utility* – the usability, usefulness or appropriateness of the visualisation (often evaluated in terms of effectiveness and efficiency), and the *soundness* – the reliability and robustness of the solution.

Despite the increasing number of sophisticated and novel infovis techniques, there is limited knowledge about the design reasoning behind many of the best practice examples (Vande Moere and Purchase, 2011). Such examples often involve a balance of *engineering design* and *creative design* processes; where the former is problem-driven and involves converging towards a single solution based on predefined requirements, while the latter involves an interplay between problem setting and solving, where many overlapping possibilities are explored before choosing a single solution (Vande Moere and Purchase, 2011; McKenna et al., 2014). These perspectives from the infovis literature compliment views expressed more broadly in Design Science Research (DSR, Hevner and Ram, 2004), where it is argued that activities from both design and natural sciences are needed in order to ensure Information Science research is both relevant and effective (March and Smith, 1995; Hevner and Ram, 2004).

In DSR, design is seen as “*a process (set of activities) and a product (artifact) – a verb and a noun*” (Hevner and Ram, 2004, pp.78). DSR can be distinguished from design in that it must produce new knowledge. It seeks out unsolved specific domain problems and produces innovative, purposeful artifacts (constructs, models, methods and instantiations) as contributions to research (Hevner and Ram, 2004). *Constructs* (vocabulary and symbols) provide the language in which problems are defined and communicated, *models* (abstractions and representations) use constructs to represent a real-world situation of the design problem, *methods* (algorithms and practices) define processes and *instantiations* (implemented and prototype systems) demonstrate feasibility by showing that constructs, models and methods can be implemented (Hevner and Ram, 2004). Hevner and Ram (2004) produce seven guidelines for ensuring a valid DSR contribution. Not only must DSR produce a viable artifact, but artifacts must be evaluated in respect to their novelty as well as the utility of solving the problem. DSR is “*fundamentally a problem-solving paradigm [which] seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis,*

design, implementation, management, and use of information systems can be effectively and efficiently accomplished (Hevner and Ram, 2004, pp.76).

The design process is described as a sequence of activities that results in an innovative artifact (Hevner and Ram, 2004). It can be the result of a system design – a logical roadmap of decisions, a user-centred design – involving users in capturing requirements, or a genius design – where designer instincts take preference (Vande Moere and Purchase, 2011). In their research, Vande Moere and Purchase (2011) introduce a model for the role of design in infovis, placing three roles of design at the centre of three different types of visualisation; practice (commercial), studies (research), and exploration (art). More recently, a design activity framework for visualisation design by McKenna et al. (2014) bridges the gap between the design models and the visualisation design decision models, with direct links to the nested model (Munzner, 2009). While there has been little explicit mention of DSR in visualisation research, there are many overlaps in the terminology, goals and evaluation activities. Indeed, well-designed and often innovative data visualisation applications are solving domain-specific problems. Visualisation is proving popular and beneficial to many domains and disciplines. In this thesis, research focuses primarily on the energy domain, specifically on visualising characteristics of household energy consumption (in the UK) but uses other scenarios to test some artifacts.

2.2.3 Energy Consumption Visualisation

Data visualisation and visual analytics offer a vast array of opportunities within the energy domain. This involves examples of network analysis and grid analytics for operators and engineers (e.g. Wong et al., 2009; Thomas and Kielman, 2009; Matuszak et al., 2013), through to the representation and analysis of domestic energy consumption patterns both for the consumer and the energy supplier. Examples are increasing with the growing quantity and quality of data becoming available from smart energy technologies.

On the consumer side, information about individual household consumption is currently reported (known as consumption feedback) in the UK through quarterly utility billing. Traditional bills are known to be vague, uninformative and do not invite householders to think about their consumption patterns (Ehrhardt-Martinez et al., 2010). Feedback rarely includes benchmarks or comparison target groups, despite research acknowledging the need for neighbourhood level comparisons to provide understandable and concrete reasoning for saving energy and encouraging discussion of energy saving techniques amongst neighbours (Räsänen et al., 2008). Research by Allcott (2011) identifies that social norms can also be used effectively in the case of energy consumption reduction and investigates this through the campaigns of OPower operating in the USA. These campaigns send letters to residents with clear graphical and textual explanation to

compare households to groups of neighbours defined as having similar characteristics. The research concludes that a small but continuous and sustained consumption reduction is achieved when this feedback continues for a long term period. A similar, yet smaller and more intensive study, used the street surface as a large visual display to allow the residents of Tidy Street in Brighton to visualise their street’s consumption in comparison to the city’s average. Results saw that neighbours could be united to fight for a common goal of reducing the street’s consumption, however, continued reduction without the visual display was only noted from a few residents and many reverted to old ways (Boucher et al., 2012). These findings suggest that the comparison of energy consumption at the neighbourhood level can offer residents a more reliable, understandable and concrete reasoning for saving energy; however, to achieve long term consumer behaviour changes in general, actions need to become integrated into daily routines and activities.

Improving user feedback is an integral part of the success of the smart meter investment: *“The main aim of smart metering is to encourage consumers to use less electricity through being better informed about their consumption patterns.”* (Firth et al., 2008, pg. 926). Advances in smart meter technology and a growing need and demand for householders to be provided with more transparent and detailed understanding of energy use within their home has seen a flurry of research evaluating alternative methods of consumer feedback (such as Ehrhardt-Martinez et al., 2010; Hargreaves et al., 2013, 2010; Darby, 2010; McCalley and Midden, 2002; Bonino et al., 2012). The traditional methods are beginning to be combined or replaced with smart energy digital displays and online profiling or usage visualisation tools. While the profiling tools and new display are shown to be more beneficial than a static energy bill (Hargreaves et al., 2010; Costanza et al., 2012), more creative and less intrusive forms of visual consumption awareness stimuli are now being investigated within the data visualisation domain (Holmes, 2007; Jonsson et al., 2010; Rodgers and Bartram, 2011; Rodgers, 2011).

While there has been much research into visualisation for the consumer, there has been little academic research investigating data analysis and visualisation from the energy supplier’s perspective. Visualisation solutions for large volumes of data could enable valuable insight into customer behaviours and habits, identify areas where consumption should be targeted for reduction, and more effectively manage supply and demand levels. This is a sensitive topic as there are privacy laws governing the use of customer data; however, the increased quantity and quality of data can provide many possibilities for personalised services and tariffs. Ellegård and Palm (with research targeting both the energy and the consumer domains: Ellegård and Palm, 2011; Palm and Ellegård, 2011) highlight how the use of visualisation helps to reveal patterns and trends in aggregated household energy data; however, this is based on diary entries of appliance use, rather

than volumes of automated data. To investigate this gap in the research, our study (explained in Chapter 3 and Goodwin et al., 2013) for E.ON, a major UK energy supplier, begins to investigate the benefits that data visualisation can bring to energy analysts by deriving value from the data emerging from new smart home technologies and identifies a number of opportunities for further research within the field. This research identifies a number of key requirements for smart home visualisation and the segmentation of energy consumer data by typical traits for market analysis. Due to the sheer number of analysis opportunities, the vastness of the design space and rapid advances in technology we are increasingly turning to creativity to help us explore the visual possibilities.

2.2.4 Creativity and Visualisation

Creativity can be viewed as a characteristic of a process, an environment, a person or a product (Rhodes, 1961 in Dean et al., 2006). Whether a (vis) design can be classed as creative or not is difficult to determine and subject to user interpretation; however, a review of the creativity research by Dean et al. (2006) reveals that most research uses a combination of the three dimensions of *appropriateness*, *novelty* and *surprise* to evaluate creativity. These dimensions of creativity are investigated and discussed in more detail in Chapter 3, where four prototype designs for energy analysts are evaluated for creativity. This research was undertaken in collaboration with academics and creativity experts within the Centre for Creativity in Professional Practice at City University London, who use known creativity techniques (e.g. Osborn, 1957; Gordon, 1960; Dean et al., 2006) within client requirements workshops to encourage engagement with users and improve requirements gathering by deliberately stimulating creativity (Maiden et al., 2004, 2007; Jones et al., 2008).

Although visualisation design can often be described as creative, the role that creativity plays in visualisation design has, until recently, been only implicit, with the creativity often stemming from the designers rather than from the users. Combining the knowledge of creativity into the domain of visualisation has begun to be explored with the explicit use of data visualisation within creativity workshops (Dove and Jones, 2013, 2014) and the amalgamation of user-centred visualisation requirements gathering methodologies (Dykes and Lloyd, 2011; Koh et al., 2011) with the use of specifically chosen techniques to explicitly stimulate creativity (Goodwin et al., 2013). The research for this TVCG publication (Goodwin et al., 2013) forms part of this thesis and is explained in detail in Chapter 3. The research forms part of the wider consideration of the design process in its broader sense (see Wood et al., 2014) and initiates a step towards a more creative visualisation design process, where a balance of aspects from

both the creative and engineering design processes are argued as necessary by the design, HCI and visualisation communities (as discussed in McKenna et al., 2014).

2.2.5 Data Visualisation Summary

The domain of visualisation is relatively new, yet has grown rapidly with the advance of computer technology and data availability. The domain covers a wide breadth of literature from design principles and perception through to data reduction and statistics. The design of visualisation is of great importance, as is the consideration of the interaction, the tasks of users, the display size and device being used. Research from the visualisation community offer models and guidelines for visualisation creation and evaluation. Although many visualisation representations can be described as creative, the academic investigation of the principles of creativity within the domain of visualisation has been limited. Creativity within data visualisation methodologies and the overlap with design science is investigated in greater detail as part of this thesis.

For the energy domain in particular, data visualisation offers many opportunities for improving the analysis of data for the network provider and energy supplier as well as improved visual representation of consumption use for the individual consumer. There are plenty of visualisation research studies analysing the network grid and investigating new and more effective visual feedback to the consumer; however, there is limited research exploring better visualisation tools for the energy companies. As academic studies indicate that energy consumption reflects socio-economic, demographic and geographical characteristics of the population, visualisation solutions for companies would allow valuable insight into customer habits, and allow for targeted services and tariffs. Further research in the area of energy consumer profiling, such as energy-based geodemographic classification, is needed in order to discover and better understand the consumer patterns and trends within the industry.

2.3 Geodemographic Profiling

As domestic energy consumption is found to correlate with demographic, socio-economic and geographical characteristics of the population in the UK, the investigation of geodemographic classification in the context of energy is highly relevant. Geodemographic classification is the clustering of small geographical areas according to social, economic and demographic characteristics of the people who live there. It is known as Area Classification, Geodemographic Classification or simply Geodemographics (Vickers et al., 2003). This section introduces the concept of geodemographics and typical use cases. This follows with domain-specific geodemographics and a discussion of whether there is a need for an energy-based geodemographic classification. The classification generation

processes are described, together with a review of the use of visualisation for improving the understanding of the process as well as the final geodemographic profiles.

2.3.1 Geodemographics

Geodemographics are based on the principle that knowing where someone lives can infer some general characteristics about their lifestyle (Harris et al., 2005). They reflect Tobler’s (1970) First Law of Geography: *“Everything is related to everything else, but near things are more related than distant things”*. The concept of grouping the population into typical traits dates back to the 19th Century, with one well-known example being Charles Booth’s 1889 Descriptive Map of London Poverty, which divided the streets of London into seven groups ranging from the ‘Lowest Class: Vicious, semi-criminal’ to the ‘Upper-middle and Upper Classes: Wealthy’ (as discussed in Harris et al., 2005).

The use of geodemographics for improving the intelligence of direct marketing, advertising campaigns and customer analysis became popular in the UK in the 1990s and continued with ever improving and competing commercial geodemographic products, the two largest of which are MOSAIC by Experian Ltd (Experian, 2014) and ACORN by CACI Ltd (CACI, 2014). Such classifications use algorithmic dimension reduction to reduce hundreds of variables to a limited number of manageable groups which reflect typical traits of the population. Classifications contain a hierarchy of clustered groups to allow for very fine profiling of the population to extremely generalised analysis. MOSAIC UK clusters 450 variables into two hierarchical steps of 15 Groups and 66 Types (Experian, 2014). Each of the profiles are named to reflect the original variables and aid the analysis. ACORN also uses hundreds of original variables to create 6 Categories, 18 Groups and 62 Types and all profiles at all levels are named individually (CACI, 2014). While the commercially available products are shown to be very useful (Ashby and Longley, 2005), a drawback is that they contain many variables from unpublished data sources (such as household and retail surveys). These, along with the methodologies used for the generation of the classification, are seen as a ‘blackbox’ to data analysts.

Following the 2001 Census, the Office of National Statistics (ONS) released a small-area geodemographic dataset known as the Output Area Classification (OAC) containing only Census variables (Vickers et al., 2003, 2005; Vickers and Rees, 2007). OAC is built on data released at the ONS statistical geography level of Output Areas (OA), which were specifically created to take into consideration the anonymity of residents, yet retain, where possible, the statistical homogeneity of the neighbourhood. OAC 2001 used 41 variables to create 7 Super Groups, 21 Groups and a further 52 (unnamed) sub groups. OAC was the UK’s first openly available geodemographic classification at this degree of detail with published methodologies and datasets. The research of OAC opened

up many possibilities for use to a wide audience; from academia and government through to business and customer analysis. There is now ample academic literature using geodemographics to cluster the UK population to better understand phenomenon that relate to demographic and geographic characteristics; for example, patterns in access to higher education (Singleton and Longley, 2009b), crime statistics (Ashby and Longley, 2005), transit services (Paez et al., 2011) and health campaigns (Petersen et al., 2011).

In mid 2014, the OAC 2001 equivalent for the 2011 Census variables, known as the 2011 Area Classification for Output Areas, was released (ONS, 2014). From here on, this will be labelled as OAC 2011 in this thesis. This classification was released by ONS and created in collaboration with research for a PhD by Gale (2014b) at UCL. The methodologies and variables of OAC 2001 were tested for optimisation and amended for OAC 2011. The new classification uses 60 instead of 41 Census variables and has 8 instead of 7 Super Groups, 26 instead of 21 Groups and 76 instead of 52 Subgroups. One of the aims of the new classification was to try to refine the groupings within London, which was known to be problematic in OAC 2001 as it was dominated by the Super Group ‘Multi-Cultural’ (Gale and Longley, 2012; Gale et al., 2014). As OAC 2011 was only released in 2014, research using this classification has not yet been reported in the academic literature; however, UCL have begun to use the new classification to test social phenomena (Longley et al., 2014).

All the aforementioned classifications have been produced as a general population profile. However, with on-going research in the area of open geodemographics there has been increased argument for specifically created domain-based classifications (Pratt et al., 2013; Singleton and Longley, 2008, 2009a,b; Gale et al., 2012): *“classification should be guided by application specific data which are intelligently sourced, validated and integrated based on existing theory applicable to specific application domains”* (Singleton and Longley, 2009a, pp.293). As research in Section 2.1 identifies a need for energy consumer profiling, this argument for specifically created domain-based classifications supports the argument for a specifically designed classification dedicated to segmenting populations based on household energy consumption.

2.3.2 Energy-based geodemographics

Despite there being a body of relevant research of multivariate statistical analysis correlating variables with energy consumption (discussed in Section 2.1), such as household income, tenure, consumption and geographical location (Druckman and Jackson, 2008; Yohanis et al., 2008; Dillahunt et al., 2009; McLoughlin et al., 2012), little research directly investigates the classification or evaluation of energy related variables with geodemographics. Druckman and Jackson (2008) compare energy consumption with

the seven 2001 OAC Super Groups, showing clear correlations with household income and property tenancy. This draws parallels with other literature (Dillahunty et al., 2009; Dillahunty and Mankoff, 2011), indicating that low-income families and tenant households have difficulties and additional barriers when it comes to reducing energy consumption.

The clustering of energy consumption variables in combination with relevant demographic variables at small-area geographies to create an energy-based geodemographic classification is ideal for investigating consumption characteristics, as not only does geodemographic research suggest human populations with similar characteristics and behaviours tend to cluster together, but energy use is found to be highly geographically and socio-economically disaggregated (Druckman and Jackson, 2008). In 2008, responding to increasing demand to be able to characterise populations based on attitudes towards green initiatives, Experian introduced a data product called GreenAware (Experian, 2008). The product allows businesses to target potential customers based on carbon footprint and includes energy variables. A case study of the use of this data by Haq and Owen (2009) demonstrates the potential for using population classifications to understand the geographical variations in energy consumption; however, like MOSAIC, it is a commercial data product and data variables and methodology are closed to investigation, and therefore difficult to rely on for academic purposes.

A particularly relevant example is a Master's thesis project by Goulvent (2012) where an open energy-based geodemographic is created using Census 2001 data at Lower Super Output Area (LSOA)¹ level, for the boroughs of Greater London. The classified dataset groups the population of Greater London into seven clusters using a total of 31 variables; 28 Census 2001 variables and 3 DECC energy consumption variables from 2009. The methodology follows the principles of OAC defined by Vickers et al. (2003; 2005; 2007). This project is a first step in proving that energy profiling can be produced and the use of open data and methodology is particularly important, as transparency of methodology and variables is paramount for the reliable use and re-creation of geodemographic classification datasets, in particular for academic research (Singleton and Longley, 2009b,a). There are, however, a number of limitations with this project. There has to-date been no validation or ground truthing of the clusters with analysts from within the industry. The classification is constrained to London boundaries so rural residential consumers are not profiled, even though rural/urban location is said to be a key variable affecting variations in consumption levels in the UK (Druckman and Jackson, 2008).

There are a number of other examples where clustering techniques have been used with energy household survey data to determine behavioural patterns and user profiles in

¹The 2nd tier Census Geography, with areas of between 400-1200 households, aggregated from Output Areas (OA): <http://bit.ly/16XucgG>

heating use (such as Guerra, 2011). Chicco et al. (2006) evaluates such techniques and methods for classifying characteristics of non-residential electricity use and adds a further use case for classifications for service providers: *“For the purpose of defining suitable tariff structures, the existing customer classifications based on the type of activity are scarcely correlated to the actual evolution of the electrical consumption and, as such, give poor information to the distribution providers”* (Chicco et al., 2006, pp.933). Since the deregulation of the electricity distribution in the UK, providers have been able to formulate dedicated tariff types (Stephenson et al., 2001). The classification of consumption characteristics will allow tariffs and services to be better targeted by consumer traits. Therefore, a new neighbourhood level geographical clustering of energy characteristics is necessary to understand complex consumption variations, allow realistic residential comparisons and enable better targeting of tariffs, services and schemes to encourage more sustainable energy use.

The clustering of areal variables (for example Census data at OA level) to form a geodemographic classification falls under what is described by McLoughlin et al. (2012) as a top-down statistical/regression approach, as it is based on aggregated datasets rather than individual households. The benefits of which are stated as being *“particularly useful [...] as they are based on real data and give a good understanding of electricity consumption patterns”* (McLoughlin et al., 2012, pg. 241). The clustering of selected variables into groups with similar characteristics is a principle data mining technique dating from the 1930s (Tryon, 1939), which is used to create geodemographics and in many other domains for the purpose of reducing large observations of data into distinguishable patterns. Large dimensions of data can also be transformed to fewer dimensions through dimension reduction such as Principal Components Analysis (PCA) (Jackson, 1991) or Multidimensional Scaling (MDS) (Cox and Cox, 2001), which often takes place prior to clustering. In the past, such techniques were costly and timely to compute; however, today the development of bespoke geodemographic classifications based on large amounts of fine detailed data is made possible through advancements in technical hardware, the availability of open-source statistical software and the wide availability of open data (Singleton and Longley, 2008).

2.3.3 Creating Geodemographics

Geodemographics are also known as Area Classifications as they are created by clustering geographical areas based on chosen variables. The process, although predominately based around the clustering, requires a series of steps, with multiple decisions to be made at each stage (Milligan and Cooper, 1987 in Vickers and Rees, 2007). The process of generating a geodemographic classification is well described in Harris et al.’s book ‘Geodemographics,

Decision	Approach A	Approach B
Number of Variables	The more input data we include the more dimensions of reality we can detect.	The more input data we include the more noise we add, therefore, masking key discriminators.
Scale of Data	Input data should ideally come from sources with national coverage.	The inclusion of large or small sample surveys can add additional information into a classification.
Normalisation	Data normalising and scaling of input data should be maximised to reduce outliers (Vickers and Rees, 2007).	Outliers should be allowed in data scaling and normalisation as they reflect reality, however, extreme values should be controlled by weighting (Harris et al., 2005).
Weighting	Input variables should not be weighted as this introduces personal bias into the construction process (Vickers and Rees, 2007).	Variable weighting provides an effective method of reducing outlier effects, thus creating more homogeneous groups (Harris et al., 2005).
PCA	Using principal component analysis is an effective means of creating classification input variables (Debenham, 2001).	Principal component analysis removes interesting dimensions of reality (Harris et al., 2005).
Algorithm Type	It is preferable to use both K-means and the Ward methods of clustering as these have a long lineage of successful commercial and academic use (Everitt, 1974).	Modern clustering methods such as partitioning around medoids or genetic algorithms provide more efficient and effective means of cluster detection from within large multidimensional datasets (Brimicombe, 2007; Brunsdon, 2006; Feng and Flowerdew, 1998).
Scale of Clusters	The clusters represented by a geodemographic classification are homogeneous across the UK.	Classification should be modelled to include regional differences (Debenham, Clarke and Stillwell, 2003).
Build Up or Down	Classifications should be built initially at the largest aggregation, then broken down into finer groups (Vickers and Rees, 2007).	Classification should be built upwards by aggregating up from the level of individuals (Harris et al., 2005).
Discrete or Fuzzy	Clusters should be represented as discrete categories.	It is more logical to represent clusters with fuzzy boundaries (Feng and Flowerdew, 1998, See and Openshaw 2001).
Open or Closed	The variable choice, weighting scheme and clustering methodology used to create a classification are commercially sensitive and cannot be made available for public inspection.	For a classification to be scientifically reproducible and safe to use, the method of construction should be open to scrutiny, preferably academically published (Longley and Singleton, 2009; Vickers and Rees, 2007).

Table 2.1: Differing priorities that guide geodemographic classification (from Singleton and Longley, 2009a, pp.293)

GIS and Neighbourhood Targeting’ (2005) together with Vickers et al.’s methodology for the creation of OAC 2001 (2005; 2007) and algorithmic validations and subsequent amendments for the 2011 OAC (Gale, 2014b).

Over the years there have been various methods for geodemographic classification and priorities change depending on the creator (the person creating the geodemographic, rather than the end user of the geodemographic classification) and the use case. These conflicting priorities which guide the decision process have been compiled by Singleton and Longley (2009a) represented in Table 2.1. The table row headings help to structure the following review.

2.3.3.1 Number of Variables

Most commercial geodemographic classifications use a large number of variables in the clustering process in order to be able to dedicate particular traits or characteristics to each cluster and create cluster names relating to the actual differences in the population. The most recent release of MOSAIC by Experian, for example, is a classification of 450 variables (Experian, 2014). On the opposite side of Table 2.1, the number of variables can be reduced to the smallest possible number to produce a general classification of the population. Any additional variables are classed as noise or duplications. This is the approach adopted by Vickers et al. (2005) for OAC 2001 and Gale (2014b) for OAC 2011, containing 41 and 60 variables, respectively.

Variables that are highly skewed or highly correlated to others can heavily bias the results and in such cases decisions are made as to whether to remove variables, transform them or use weighting (Harris et al., 2005). Variables were also rejected from OAC 2001 and 2011 for having an “uninteresting geographic distribution” or for lack of consistency over time (Vickers and Rees, 2007). In terms of geography, ethnic group variables were examined for inclusion in OAC 2001. The distribution of White and Chinese populations were more homogeneous and perceived to be poor indicators for a classification, compared to other ethnic groups such as Black Caribbean, Black African, Black Other, Indian, Pakistani and Bangladeshi, which were all included in OAC 2001 (Vickers, 2006). Gale (2014b) did not automatically discount these variables for OAC 2011, noting that the results of the 2011 Census show that England and Wales are becoming more ethnically diverse, with the White ethnic group accounting for 86% of the population, down from 91.3% in 2001. Geographic variation was one of four variable reduction indicators for OAC 2011 together with Pearson’s r correlation analysis, the within-cluster sum of squares (WCSS) analysis and skewness (Gale, 2014b).

The consistency of data over time is also particularly necessary for OAC in comparison to commercial geodemographics, as the Census data soon becomes

out-of-date with considerable population and urban change occurring during the ten-year Census cycle (Gale and Longley, 2013). Commercial products, on the other hand, are built on hundreds of household surveys which are updated regularly and include current year estimates to key Census variables (Experian, 2009). This is a particular challenge for open geodemographics such as OAC, as the future of the UK Census in its current form is unknown (Martin, 2006). Open data is therefore beginning to be investigated as an alternative data source. The Economic and Social Research Council are funding the ‘Using Secondary Data to Measure, Monitor and Visualise Spatio-Temporal Uncertainties in Geodemographics’ project at the University of Liverpool, which forms an initial step in the direction of integrating other sources of open data into open geodemographics (Gale, 2014b). Commercial geodemographics are also moving away from relying heavily on the Census; CACI have recently renewed their methodology for their ACORN product to reflect these changes and in some areas of the UK no Census variables are needed for the classification (CACI, 2014). Irrespective of the total number of variables, it is clear that careful attention must be given to the data input comparability, the avoidance of duplication of existing information, updatability, robustness and scale (Harris et al., 2005).

2.3.3.2 Scale of Data

There are multiple definitions for scale. Two dimensions relevant for this research are defined as *Scale Resolution* (SR) – the level of aggregation of the data – and *Scale Extent* (SE) – the geographical extent of the data (Lamand and Quattrochi, 1992; Turkay et al., 2014). These are explained in relation to spatial scale, through illustrations in Fig. 2.2.

Traditionally, data for geodemographics was sourced for 100% of the extent of the output area, which was most often national coverage (Harris et al., 2005). This is the method that both OAC 2001 and 2011 use; however, commercial classifications incorporate other data sources such as household surveys. These surveys are often at finer resolutions and may not be representative of the whole study area (extent) (Harris et al., 2005). The SE needs to also be considered when creating a geodemographic classification, as clustering data covering the whole of the UK (e.g. OAC 2001 and 2011) produces very different results than for a reduced extent; for example a classification for the extent of London (e.g. LOAC) (Petersen et al., 2011; Gale et al., 2012).

In terms of SR the nature of the summaries used to describe areas at each scale can vary, as can the relationships between them. While the data priorities in Table 2.1 do not refer to the use of data at different resolutions it is evident that OAC 2001 and 2011 use only one geographical resolution (Output Area) for all data variables across the UK, while commercial products, such as MOSAIC, combine data at different resolutions depending

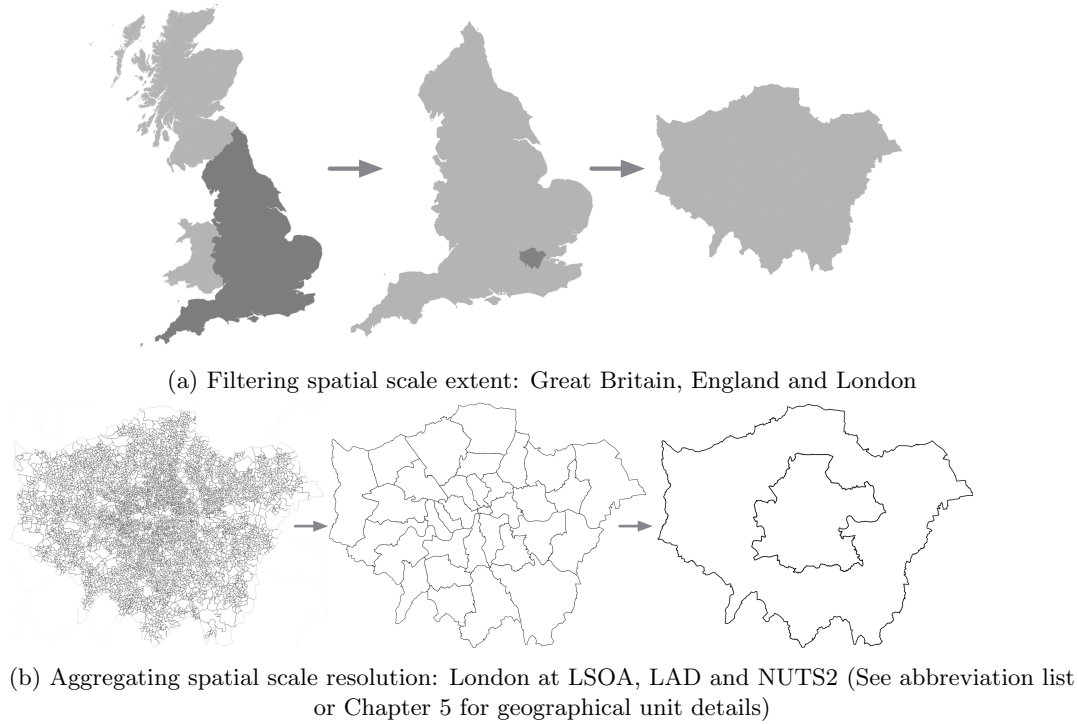


Figure 2.2: Spatial scale extent and resolution examples. Contains National Statistics data ©Crown copyright and database right 2015, and OS Data ©Crown Copyright and Database Right 2014

on the data source available (Harris et al., 2005). The incorporation of data at different resolutions into the geodemographic classification process may help to differentiate social phenomena occurring at different scales (see Section 2.3.3.4). However, as commercial products do not divulge variable sources or their specific methodologies there is limited academic research in this area at present. Nevertheless, with the unknown future of the Census (Martin, 2006) and the recent influx of open data sources at different scales (both SR and SE) recent research in open geodemographics discusses the need to adapt the traditional methods in the future (Gale, 2014b).

While it is not yet clear how beneficial the use of data at different scales is to geodemographic classification, the aggregation of the data can influence clustering results, with the effects of “badly behaved distributions” being more acute at small spatial scales, where the likelihood of extreme values increase (Vickers and Rees, 2007, pp.384). Data aggregation is well known to greatly affect outliers, which can lead to the *modifiable unit problem* (MAUP) (Openshaw and Taylor, 1984) and result in *ecological fallacy* (Monmonier, 1996) – the fallacy of making inferences about individuals from aggregate data. MAUP comprises of two parts; the scale problem – relating to the size of the zones – and the zone problem – relating to the shape of the zones (Lloyd, 2014). The complexities of MAUP have been studied in academic literature (e.g. Openshaw and Taylor, 1984; Fotheringham and Wong, 1991) and are seen as a fundamental problem which impacts all studies of spatially aggregated data (Lloyd, 2014). Prior to 2001 the

smallest spatial unit of output geography for the UK Census used arbitrary units designed for statistical collection rather than designed for optimal statistical output. Areas were mixed in terms of size, shape and social homogeneity and therefore MAUP was a real concern. Since 2001 the Census data has been released at OA level and data at this level were used to create OAC 2001 and 2011. OAs are geographical units specifically designed using an automated zoning procedure to produce an optimal arrangement in terms of social homogeneity, shape and size and thus reduce the effects associated with aggregation (Martin, 2000). In addition to OAs, two higher tiers of Census level geographies are also created in a similar manner: Lower Level Super Output Areas (LSOA) and Middle Level Super Output Areas (MSOA).

Geodemographics are produced at many aggregations of geography and the results vary considerably at each level; for example, area classifications for Census 2001 are available at 5 levels: Health Areas, Local Authority, Wards, LSOA and OA². Monmonier (1996) states that it is important to choose the most appropriate scale for the specific purpose of analysis. Vickers (2006) notes that it is especially important for area classifications, as regions are not only grouped based on data but given names and descriptions based on the values. Therefore, using an inappropriate scale for area classification not only produces poor or false analysis, but can cause offence. A drawback of OAC, in comparison to commercially available geodemographics, could be the scale of data resolution, as commercial products usually use smaller postcode units. Despite new methodologies, commercial geodemographics are often still heavily based on Census variables and data is refined to the postcode level with the addition of many other data sources and disaggregation calculations; for example MOSAIC 2009 contained 440 variables, 38% of which are based on Census variables (Experian, 2009). Singleton (2007) argues the benefit of the use of OAC for neighbourhood analysis by investigating the degree of detail obtained from using MOSAIC classification instead of OAC. The research concludes that there is little heterogeneity between postcodes within an OA which supports the argument for the use of OA data for neighbourhood analysis.

Changes to geographical regions over time are also challenging. In the case of the most recent Census (2011), the OAs, which were first built for the release of the 2001 Census, were kept as near to 2001 boundaries as possible; however, a small proportion had to be adapted to reflect administration changes, population growth or the statistical homogeneity of areas (Cockings et al., 2011; Tait, 2012b). In terms of the three tier geographical output for the 2011 Census for England and Wales, there are now 181,408 OAs of which 2.1% changed between 2001 and 2011, 34,753 LSOAs of which 2.5% changed and 7,201 MSOAs of which 2.1% changed (Tait, 2012b). While this seems a small percentage

²Official ONS Area Classifications 2001 available at: <http://bit.ly/OACGeo>

with respect to the whole country, these changes affect the densely populated urban areas. Regions were merged and/or split. Comparison across two geographical units, which are not hierarchical, is difficult, unreliable and an area of concern for merging datasets. This is discussed further in Section 5.4.1, where two datasets using Census geographies from the two years are combined.

In general, when creating geodemographics, the scale of data resolution and extent must be considered carefully as the results at each level can differ substantially. As population characteristics and behaviours vary at different scales of resolution, there is also a need to capture the scale at which the phenomenon occurs. Determining the most appropriate scale of analysis for each variable or the collection of variables for clustering would be useful for the justification of the classification. The compatibility of the geographies across time, as well as the method used to define the geographical areas, are also important considerations for the accuracy of the datasets.

2.3.3.3 Normalisation, Transformation, Weighting and PCA

Vickers et al. (2007) report that classification variables must be like-for-like in terms of their data range to avoid bias in the resulting clusters. In data mining or machine learning it is common to normalise values when using clustering methods in order for them to be in the same range (Harrington, 2012). Outliers, data scaling and normalisation are dealt with in different ways in geodemographic research (Singleton and Longley, 2009a; Vickers et al., 2005; Harris et al., 2005). OAC 2001 and 2011 use a range standardisation to ensure comparability and remove outliers, while other classifications often use the z-score (Harris et al., 2005). Prior to standardisation, each OAC variable is translated to a relevant percentage or ratio.

Vickers et al. (2005) demonstrate that there is additional benefit to transforming the data to a log scale prior to the range standardisation process, in order to reduce the effect of the outliers on the clusters even further. The variable ‘population density’ was highlighted as being a particular problem in OAC prior to this transformation. This double standardisation was seen as necessary for OAC 2001, as logging the data reduces the likelihood of a highly skewed distribution within a variable. The benefit of using the logarithmic value in combination with the range standardisation is evaluated by Vickers et al. (2005) by comparing the mean value for each variable prior to and after introducing the log value into the equation. The use of the log transformation in OAC 2001 is described by Vickers et al.: *“to reduce the effect of large gaps between variable values, which were typically found at the higher end of the range of values. The log transformation of the data squashes the ends of the data series and expands the middle”* (Vickers et al., 2005, pp.43).

Gale (2014b) uses an alternative transformation method (Inverse Hyperbolic Sine) for OAC 2011, having tested three alternative methods for best fit. Not only were three normalisation methods tested but three algorithmic methods for each stage of data preparation (Percentages, Index Scores and Mean Difference), normalisation (Log, Box Cox and Inverse Hyperbolic Sine) and standardisation (Range, Z-Score and Inter-Decile Range) were tested, resulting in 27 unique datasets. These datasets were compared statistically by analysing the skewness of the datasets together with the results of clustering the datasets into clusters of 6, 7 and 8. These clusters were then tested for low levels of cluster homogeneity. From the 27, four datasets were investigated further through interpreting the geographical distribution of the resulting clusters and finally the ‘Percentages, Inverse Hyperbolic Sine, Range-Scale’ dataset was selected for the final classification (Gale, 2014b).

The method of principal components analysis (PCA) is common in other examples of clustering for variable reduction (Harris et al., 2005). PCA is an algorithmic methodology which trims the dataset to remove redundant information which is not primary to the clustering, and thus removes outliers and repeating information. This is not recommended for geodemographics, as it can *“blur rather than clarify fine distinctions between cluster types”* and the results are no longer intuitive (Harris et al., 2005, pp.157). As the final stages of the classification involve naming the profile in respect to the original variables, using the results of PCA analysis would make this particularly difficult. In geodemographics, transformation or weighting are often used instead to reduce the emphasis of skewed or overlapping (but still seen as important) variables. Despite continued use for OAC, the transformation of variables is questioned in the literature. Harris et al. (2005) state that variable outliers are important to the geodemographic process and thus transformation is not necessary. In this case, variable weighting is considered important only where extreme values need to be controlled. Harris et al. (2005) explain that variable weighting can be determined through correlation analysis, the creation of minimum spanning trees, PCA or prior experience.

2.3.3.4 Clustering Algorithm and Method

There are numerous clustering algorithms available for different purposes, with k-means being one of the most common in the area of geodemographics and customer segmentation (Harris et al., 2005; Linoff, 2011; Vickers et al., 2005; Gale, 2014b). K-means is suitable for geodemographics as it can handle large numbers of variables and is capable of handling outliers, common in demographic data, better than other clustering algorithms. The k-means algorithm aims to partition observations by k number of clusters, with an objective of minimising the given variance within a cluster.

The method uses an iterative relocation algorithm, where a case is moved from one cluster to another and repeatedly refined until a stable cluster is reached and no further moves can occur. Performance is assessed using the error sum of squares (Vickers and Rees, 2007). The sample size and scale of the data is important, particularly for k-means where there must be a high sample size for all areas (Vickers and Rees, 2007).

Clustering methods in general are known to be sensitive to small perturbations and outliers as well as differences in scale (Ledolter, 2013). According to Harris et al. (2005), Experian uses two types of clustering methods in their approach with the stepwise approach and then the allocation method (k-means). Due to the differing scales of the data sources, Experian uses concentric circles to include more and more regional content at each step, and thus allows for differentiation of similar demographics in rural and urban settings; an example of locally clustered geodemographics (Harris et al., 2005). The concept of geographically weighted geodemographics is beginning to be investigated in detail for open-geodemographics (Adnan et al., 2013).

In terms of ‘Build Up or Down’ (Row 8, Table 2.1), Experian use a clustering hierarchy from the bottom up by clustering the data into types and then combining these types into groups. In OAC the hierarchy is produced in the opposite way, where Super Groups are clustered first and then each is sub-clustered to produce the Groups and Sub Groups. OAC 2011 follows the same approach as OAC 2001, as user consultation identified a preference for a top-down hierarchical structure, with a similar number of clusters if possible (ONS, 2014).

Tools which have been produced to aid the generation of geodemographics such as GeodemCreator, a UCL PhD project by Adnan (2011) and gd a package for the R Project (open source statistical software: <http://www.r-project.org>) created by Alex Singleton (2012). Both tools use k-means for the basis of the clusters, allowing the user to define the number of clusters, re-run the calculations to produce stable results and investigate the results. Other suitable clustering algorithms are ward (Vickers and Rees, 2007), hierarchical and k-medoids (Zhao, 2012) and as is shown in Table 2.1 (Row 9) there are also options for the production of fuzzy clusters, which may better represent the population. While the aforementioned examples use discrete boundaries for the clusters, there is also research that investigates the use of fuzzy boundaries within geodemographics (See and Openshaw, 2001; Son et al., 2012; Fisher et al., 2014).

2.3.3.5 Openness, Validation and Naming

As often discussed in the literature (Longley and Singleton, 2009; Vickers and Rees, 2007; Gale et al., 2012; Gale, 2014b), there is a need for geodemographics to be open in order to be scientifically reproducible and safe to use, especially when policies are being made

on data insights. The results of the clustering can be validated to test whether there is significant structure in the data, that each cluster is representative of the population, that clusters are not overly sensitive to the use of one variable and that the clusters are indeed stable (Vickers and Rees, 2007). The naming of the clusters in geodemographics is also an important stage, although often contentious, with different users wanting different types of labels and some groups being mindful of political correctness (Harris et al., 2005). This usually involves creating labels and pen-portraits with comprehensive descriptions of each class. OAC 2001 and 2011 names and pen-portraits were produced with public involvement and consultation.

In general the whole process of generating a geodemographic classification is subject to decisions at each stage and the experience of the creator plays an important role. Harris et al. describe the process as “*as much an Art as a Science*” (2005, pp. 159). The opportunity to interact, try various schemes and strategies, learn from past experiences and create the optimal solution for your data is all deemed as knowledge discovery.

2.3.3.6 Creating domain-specific Geodemographics

In addition to the process of generating a geodemographic classification, Singleton and Longley (2009b) demonstrate an alternative method to create a domain-specific geodemographic based on existing geodemographics, in this case OAC 2001. OAC variables are first re-clustered to a finer level of 146 clusters and then the results are combined with higher education variables to create a new higher education-based geodemographic classification. Two possible methods for domain-specific classification have therefore been identified:

1. Adapt the existing OAC by firstly re-clustering the OAC variables to a finer number of clusters and then adding the domain-specific variables and clustering a second time to build a profile of a smaller number of clusters (following the methodology of Singleton and Longley, 2009b)
2. Re-cluster a classification from scratch (e.g. from OA level upwards) based on the experience of creating OAC (2001 or 2011), with selected OAC variables and additional domain-specific variables (similar to the process in Adnan, 2011; Goulvent, 2012)

Creating a well-defined, usable and justifiable geodemographic classification is time consuming, and without proper documentation and justification, the results are subject to misinterpretation. Although improvements are being made in the area, for example with the GeodemCreator tool allowing bespoke domain and local geodemographics to be created on the fly (Adnan, 2011), the gd package for R (both investigated in more detail in

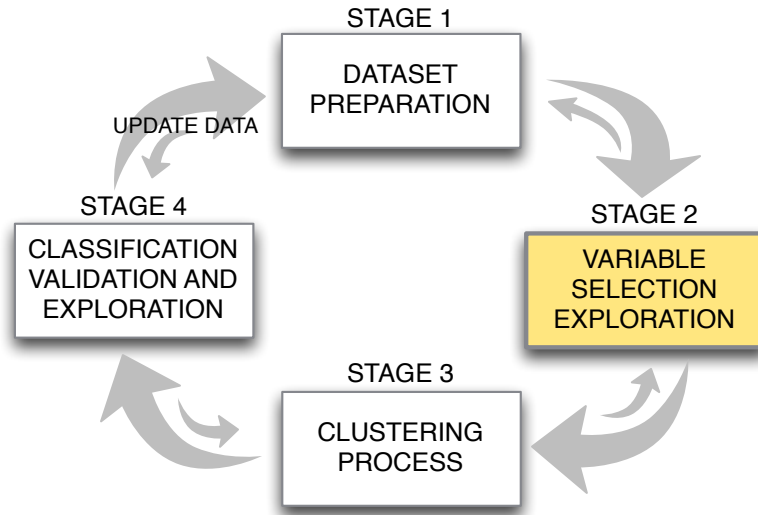


Figure 2.3: Simplified four stage circular process for generating a geodemographic classification. The research in this thesis focuses particularly on Stage 2.

Section 4.2.1) and most recently the release of the code for the creation of OAC 2011 (Gale, 2014b)³. The process of generating domain-specific or local geodemographics is currently mainly limited to academic experts in the domain (e.g. Singleton and Longley, 2009b; Goulvent, 2012; Gale et al., 2012).

2.3.4 A Four Stage Process

The previous sections demonstrate that the process of generating a new geodemographic classification is time-consuming and complex. There are different approaches which will result in different outputs, each are dependant on the creator, the variables and the use case. For the purpose of this research, the evaluation of the literature has lead to the combination of these processes into four inter-dependent stages (shown in Fig. 2.3). The four stages are explained as follows:

1. **Dataset Preparation:** involves the investigation and combination of potential variables from various data sources. Variables from the different data sources are usually combined into one table and the standardisation of the spatial scale of resolution and extent is important to ensure the comparable of the variables. The data collection period should also be comparable, to avoid uncertainty in the patterns. At this stage preliminary decisions on groupings of variables are made, for example % of population aged 0-4 or aged 0-9, although this decision can be altered at a later stage if deemed to be beneficial to the clustering (hence, the arrow back from the later stages). Variables are linked to a base count and converted to percentages or ratios and the variables standardised (e.g. Range Standardisation was used for OAC 2001 and 2011) to ensure they are comparable.

³OAC 2011 R code is available at: <https://github.com/geogale/2011OAC/>

At this stage, careful attention is given to the data source including scale, updatability, reliability, robustness and comparability.

2. **Variable Selection Exploration:** involves the in-depth and careful investigation and comparison of the candidate variables. To avoid causing bias to the clustering process each variable should be independent, of near-normal distribution and have little or no pairwise correlation. Whether a variable is geographically distributed is also of interest, as variables with limited geographical distribution will have little or potentially negative impact on the classification. Multivariate comparison is key to this process and depending on whether many variables are used (as in commercial products) or as few as possible (as in OAC) decisions are made on whether to merge, drop, split or to (go back to stage 1 and) transform the variables to reduce the skewness and remove heavy outliers, for example by a logarithmic scale such as OAC 2001 or the Inverse Hyperbolic Sine for OAC 2011.
3. **The Clustering Process:** the clustering algorithm is chosen (most often k-means is used for geodemographics, as discussed in Section 2.3.3.4), appropriate class numbers are identified and data values are clustered. Certain variables may prove to dominate these clusters and variable weighting can be implemented to reduce the effect of variables on the clusters. The cluster hierarchy is also relevant at this stage, i.e. whether the clusters are created bottom up or top down (See Section 2.3.3.4).
4. **Cluster Validation and Exploration:** The final stage of the process involves interpretation, validation, replication and cluster naming. Unlike other forms of data classification or clustering, for geodemographics it is necessary to investigate the clusters manually in detail, sometimes merging or splitting clusters in order to create a profile which can be named based on the characteristics of the population it describes. As mentioned in Section 2.3.3.5, clusters can be tested in various ways, such as testing the sensitivity of removing each variable, cross-validation to ensure they are representative of the population and replication to ensure that the cluster is stable (Vickers and Rees, 2007). Naming the variables often involves visual representations of the clusters to help create names and pen-portraits.

OAC methodology states that variables should stay consistent over time (Vickers and Rees, 2007), but this may not be the case for commercial geodemographics, which are up-dated on a more regular basis (CACI, 2014; Experian, 2014). For the creation of profiles where data is updated regularly the process will continue in the circle, as shown in Fig. 2.3. The geodemographic literature (Vickers et al., 2005; Vickers, 2006; Goulvent, 2012; Gale, 2014b) and personal communication from those building open

geodemographic classifications (Gale, 2014a) confirm that it is not a linear process, but usually involves going back and forth at each stage to ‘iron out’ issues (e.g. most notably by removing, transforming or changing the weighting of variables) prior to establishing the final classification, hence the bidirectional arrows in the diagram (Fig. 2.3). These four stages (and diagram) are referred to throughout this thesis; the use of interactive visualisation within the process is investigated in Section 4.2 where Stage 2 is identified as needing further research and forms the context for Chapters 5-7.

2.3.5 Visualising Geodemographics

While the use of clustering to create geodemographics radically reduces data volumes and enables trends and clusters in large datasets to be identified with greater ease, they are often misinterpreted (Harris et al., 2005). One of the major drawbacks of geodemographic classification is the uncertainty resulting from the cluster decision process, as the diversity and variance within the clusters can be large and overlapping or multiple cluster ownership often occurs, yet this is not necessarily reflected in the outputs (Vickers and Rees, 2007; Harris et al., 2005).

Research in the infovis domain show that visualisation is beneficial when used in collaboration with data mining techniques such as clustering (Shneiderman, 2002). In order to aid the users interpretation of the clustered dataset, visualisation has been shown to be useful at various stages of the process, in particular Stages 3 and 4; for example, the cluster decision process (Cao et al., 2011; Choo et al., 2010), cluster verification – where visual representations such as radial diagrams are often used to distinguish dominate variables and name profiles (Harris et al., 2005) – and for the representation or reporting of the final classification (Slingsby et al., 2010b, 2011). The research by Slingsby et al. (2010b; 2011) enables the user to investigate the uncertainties in detail using visual variables, such as colour, to determine the degree of dominance of a cluster. More recent work investigates the uncertainty through a fuzziness parameter for fuzzy group membership (Slingsby et al., 2014b; Fisher et al., 2014). The findings reveal that the visualisation aids user’s understanding of such complex classified datasets and that well-designed and novel visualisation methods can be used effectively to overcome some of the problems associated with the kinds of visualisation that are more routinely used to demonstrate geodemographics, such as thematic maps.

The initial stage (Stage 1) of the classification process is the data source investigation, which lends itself least to visualisation as it is reliant on searching for external data sources and combining variables ready for comparison. However, the data merging could benefit from visual representation, as missing data or overlapping regions could be flagged to the user for further inspection. The next stage of the process (Stage

2), where variables are compared for correlations and skewed distributions, known in clustering as feature selection, is popular in visualisation research (e.g. Yang et al., 2004; Seo and Shneiderman, 2005; May et al., 2010, 2011; Krause et al., 2014); however, these are limited to other examples of classification and are not specific to geodemographics. As geography is an additional dimension to consider when choosing variables for geodemographics, there is need for more specialised visualisation tools specific to geodemographic variable selection. Geography in OAC 2011 was investigated through the use of a qualitative measure, rather than a visual representation. Visualisation was, however, used to evaluate variable skewness through small multiple distribution plots and matrices were used to compare variable correlations (Gale, 2014b).

While the use of PCA for variable reduction is not recommended in geodemographic classification (as discussed in Section 2.3.3.3) it is common in other forms of clustering, and visualisation literature is available (such as Ingram et al., 2010; Turkay et al., 2012). As discussed in Section 2.3.3.3, data transformation was performed in OAC 2001 and 2011 to transform the heavily skewed distributions. The transformation of variables, either statistically or through PCA (Jain et al., 1999), can be problematic, as the reasons for exactly how and why the variables are transformed during this process is difficult to interpret (Harris et al., 2005). Transparency of the process through visual means, allowing the users to clearly see how the transformations affect the original variables and the clusters, can enable the process to be more comprehensible (Seo and Shneiderman, 2005).

It is evident that the process of generating a geodemographic classification contains many uncertainties and sensitivities relating to decisions made at each of the stages of the process. Harris et al. (2005) describe the process as an inherently creative process, much like visualisation was defined and described in Section 2.2. There is great potential for the use of visualisation, in particular visual analytics, to aid the generation of geodemographics, as they are described as tools for engaging users with the data and for adding transparency into machine learning and statistical processes. A visual tool to progress through the four stages of the process of generating a geodemographic classification, building on current tools which aid the creation of domain-specific geodemographics (e.g. Adnan et al., 2013; Singleton, 2012), could benefit the creator. The use of multiple views, for example an overview of the variables, as well as detail-on-demand, would be beneficial, as would the ability to store stages and document the process for aiding the final validation stage and allowing for reproducibility (Shneiderman, 1996). A visual process for aiding the creation of an energy-specific classification process also follows the academic call for geodemographics to be brought into the current data and technical era and follow more domain-specific,

problem-centred approaches utilising the advances in visualisation and data exploration techniques (Singleton and Longley, 2009a). Incorporating visualisation in the process may also help to bridge the gap between the domain experts and the middle-ground users, who may not have the in-depth mathematical knowledge of the process currently necessary, but have a need to create domain-specific or local geodemographics, as targeted by DimStiller (Ingram et al., 2010) for visualising dimension reduction.

2.3.6 Geodemographics Summary

This section introduced the concept of geodemographics. The geographical, socio-economic and demographic correlations with energy consumption indicate that an energy-based geodemographic classification would be beneficial for profiling energy users of the country. While such an energy-specific classification follows a recent call for domain-specific geodemographics, the generation process is complex, time-consuming and intensive, and decisions made at each stage can affect the results. Knowledge of the statistical algorithms used, as well as the variables themselves, are necessary for a successful process. The process of generating a domain-specific geodemographic classification is simplified into four interdependent stages for the context of this research. The discussion leads to the argument that each stage of the process would benefit from specifically designed visual tools to engage the creator, aid understanding of the algorithms and open the process to a wider user group. There are a number of examples where visualisation has been used with geodemographics to highlight the uncertainties in the process and to visualise the clusters and profiles. The variable selection stage (Stage 2) remains open to investigation with many useful examples, yet limited in the context of generating geodemographic classifications, where variable scale and geographical variation play a key part in the final decision process.

2.4 Multivariate Comparison

As explained in the previous section, the selection of variables through comparison is an important part of building the classification, as the variables selected should be independent, of near-normal distribution and have little or no correlation to one another. The variable selection process, also known in clustering as feature selection (Jain et al., 1999), is time-intensive and subject to user interpretation. Vickers (2006, pp.86) states that for geodemographics the process is *“far from straightforward and cause of much debate”*.

To initialise the process, Vickers (2006) states that PCA can be used to determine which of the initial variables have the strongest influence on the dataset. This is, however, not an official reason for rejecting variables in OAC 2001. These reasons are: highly correlated variables, variables with badly behaved distributions, composite

variables, variables without national coverage or geographic consistency, vague or uncertain variables, variables with uninteresting geographic distributions, and variables with reduced consistency over the lifetime of the classification (Vickers et al., 2005; Vickers and Rees, 2007). These criteria indicate that the decisions are based on the knowledge of the variable and data sources themselves (e.g. consistency of the variable over time) together with three main aspects of variable comparison:

- **Distribution:** heavily skewed variables are investigated and subsequently removed, transformed or weighting is heavily reduced in order to reduce their effect on the clustering;
- **Correlation:** strongly correlated variables are investigated concurrently to reveal duplicate, similar or composite variables;
- **Geography:** variables with geographically variant distributions (and full coverage) are preferred to ensure the geographical spread of the clustered profiles.

Following the recent release of the methodology for OAC 2011, Gale (2014b) also uses within-cluster sum of squares (WCSS) to identify which of the 167 variables have the greatest impact on the classification. Here, the top 10 negatively impacting variables for each of the 27 datasets were identified and further investigated using the three above comparison measures. This can be seen as an alternative to the use of PCA to determine strength of variable influence, as used by Vickers (2006).

The process relies fundamentally on multivariate statistics as a means to evaluate, rank and compare the variables. Multivariate statistical analysis is inherently more complex than uni- or bi-variate statistical analysis due to the need to understand the relationships between many variables: “*The human mind is overwhelmed by the sheer bulk of the data*” (Johnson and Wichern, 2007, pp.1). Due to the volume of data and complexities in comparison, multivariate statistical analysis goes hand-in-hand with graphical techniques to visualise and aid the analysis process. It is noted in statistics, that plotting is often neglected (Johnson and Wichern, 2007); however, in the visualisation community, examples of multivariate comparison using the abstract tasks of understanding distributions, outliers and correlations are extremely common (Munzner, 2014). Increasingly, data visualisation tools are being built to support such comparisons, allowing users to examine each object individually. The design of the visualisation is particularly important in the case of visual comparison. Gleicher et al. (2011) describe comparative visualisation as being a mixture of juxtaposition, superposition and explicit encodings; where *juxtaposition* represents each object separately, next to each other in time or space, *superposition* presents multiple objects on top of one another in the same

coordinate system and finally *explicit encodings* refers to the visual encoding of the relationship provided through the computation of the relationships between objects.

The following sections discuss relevant multivariate comparison statistical techniques related to distribution, correlation and geographical variation together with examples of their representation from the visualisation community.

2.4.1 Distribution

The distribution of a data variable describes how often each value appears (Downey, 2011). The most common graphical representation of distribution is a histogram, which shows the frequency of each value as a count in different bins. They are effective at visually representing distribution as they immediately make the model value apparent; outliers are easy to identify and the shape of the histogram determines the type of distribution: normally distributed (a bell curve), asymmetrical (as either positively or negatively skewed), or bi- or multi-modal (with more than one peak). In order to describe a distribution, its central value (e.g. the mean or median) is identified, together with a measure of the variance of the data (e.g. standard deviation) (Few, 2004; Reimann et al., 2008). Data quartiles or percentiles of the distribution may also be of interest. In terms of statistical summaries there are also measurements such as kurtosis or skewness. Kurtosis provides an expression of the curvature and skewness measures the symmetry of the distribution, where the sign of a value indicates the degree of negative or positive skew of the distribution. The skewness of a variable should be in the range of ± 2 , else the skewness of a variable can be classed as extreme (Reimann et al., 2008). These are frequently used in data comparison, although the presence of outliers or multi-modal distributions is known to bias results. When choosing statistical measures for multivariate analysis, it is important to consider more reliable and robust methods of statistics, which are resistant to such outliers and variance in the data (Filzmoser et al. 2008 in Turkay, 2013).

In terms of visualisation, the histogram on its own is often not sufficient and many other forms of visual representation can be applied. The boxplot for instance *“is one of the most informative graphics for displaying a data distribution”* (Reimann et al., 2008, pp.41) and is also known as the Tukey plot (Tukey, 1977) or the box-and-whisker diagram. These can be plotted in juxtaposition and are therefore ideal for multivariate comparison. Edward Tufte (1991) also introduced a minimalistic version by maximising the data-ink ratio, making it even easier to combine multiple plots on one screen. Another minimalistic alternative is distribution plots, which indicate the shape of the histogram through a line rather than individual bars. Gale (2014b) uses small multiple distribution plots to identify the skewness of variables for OAC 2011 (see Fig. 2.4). Other typical graphics for

representing distribution include density traces, one-dimensional scatterplots, frequency polygons, ECDF or CP plots (Few, 2004; Reimann et al., 2008).

As explained in Section 2.3.3, both OAC 2001 and 2011 used variable transformation to normalise skewed distributions. This transformation process is run on all variables, rather than just the few isolated skewed examples. Statistical literature advises the transformation of scale of all the variables in the same manner, as there are empirical and scientific implications when transforming data (Field, 2013). The use of animated transitions or parallel plots can aid the transparency and comprehension of the process and allow users and creators to clearly see how the transformations affect the original variables and the clusters (Seo and Shneiderman, 2005). Introducing thresholds, an alternative to transforming variables to remove outliers, is also often implemented on graphical representations, e.g. at the 95 or 98% range (Tufte, 1991).

2.4.2 Correlation

Correlation estimates the extent of the relationship between any pair of variables. The covariance is the measure of how much two variables change together. This is represented as any number where only the sign (+ or -) is informative and the magnitude of the relationship is difficult to comprehend. Instead, the normalised version of the covariance, the correlation coefficient, is often used. The coefficient shows the strength of the linear relation between variables by its magnitude which ranges from between +1 and -1. To interpret the scale ± 1 represents a perfect (positive or negative) 1:1 relationship, while 0 indicates there is no relationship between the two variables. The three most widely used calculations of correlation coefficient are the Pearson, Spearman and Kendall (Galton, 1890, Spearman, 1904, Kendall, 1938 in Reimann et al., 2008). The Pearson method is often used as default and is ideal when data follows a normal distribution; however, it is very sensitive to data outliers (Reimann et al., 2008). The Pearson method is also sensitive to aggregation, where information can be lost through reducing the number of scaling points (Martin, 1978). This is particularly significant when both variables have fewer than 10 points. There are, however, alternative correlation measures which can correct for scaling problems (as discussed in Martin, 1978).

In terms of visual representation, the relationship is often presented as a scatterplot, with each variable on each axis. The scatterplot allows users to not only judge the correlation, but provides an overview of the variable and characteristics such as the variable distribution, outliers and extreme values. Regression lines can also be calculated and superimposed. Multiple scatterplots can be compared concurrently when arranged in small multiples, whereby each row and each column represent a variable and cell positions within the matrix determine each variable combination. These multiple

2.4. MULTIVARIATE COMPARISON

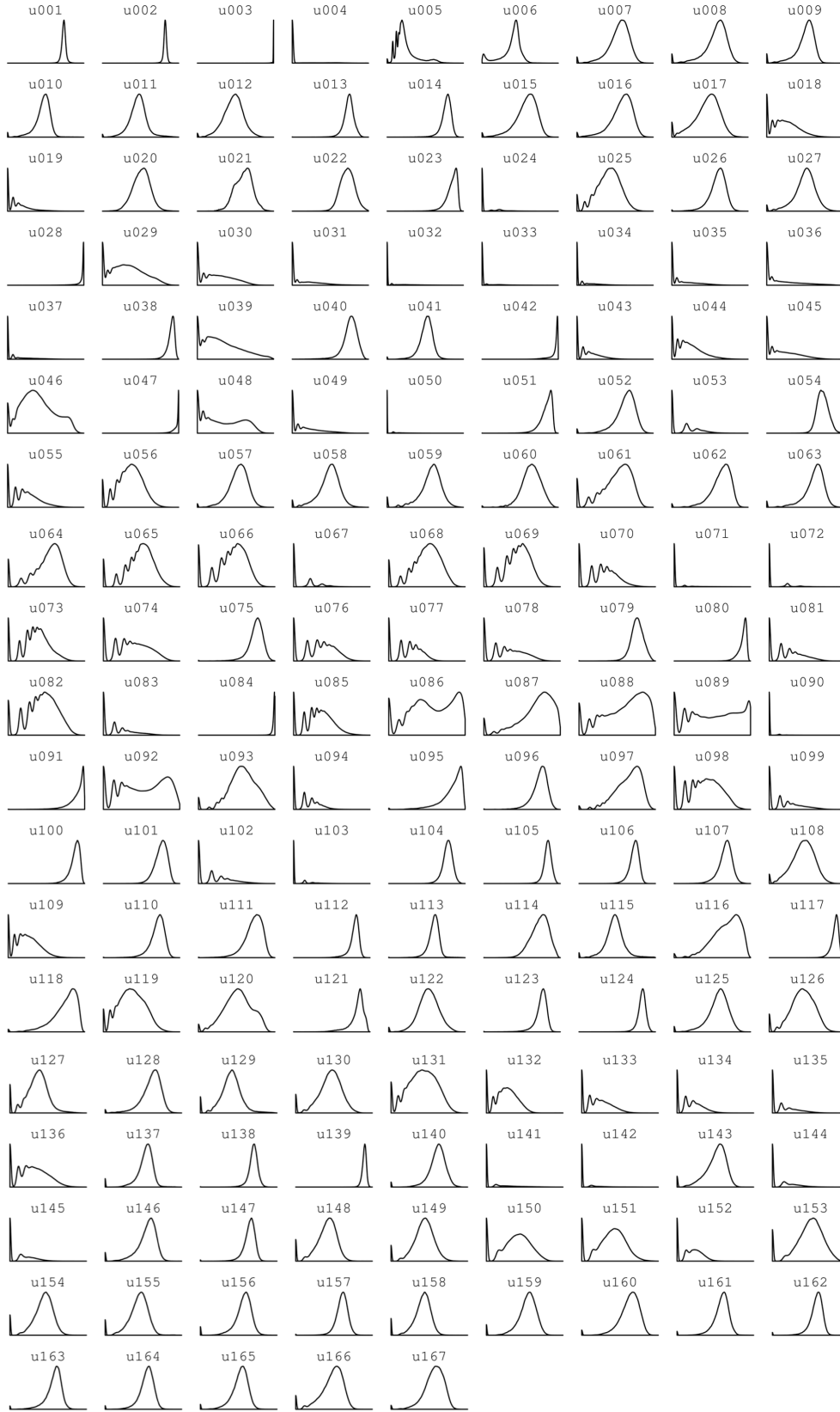


Figure 2.4: Distribution plots for the 167 initial variables for OAC 2011 to represent variable skewness from Gale (2014b, pp.469-471). See Appendix B.8 for the variable names.

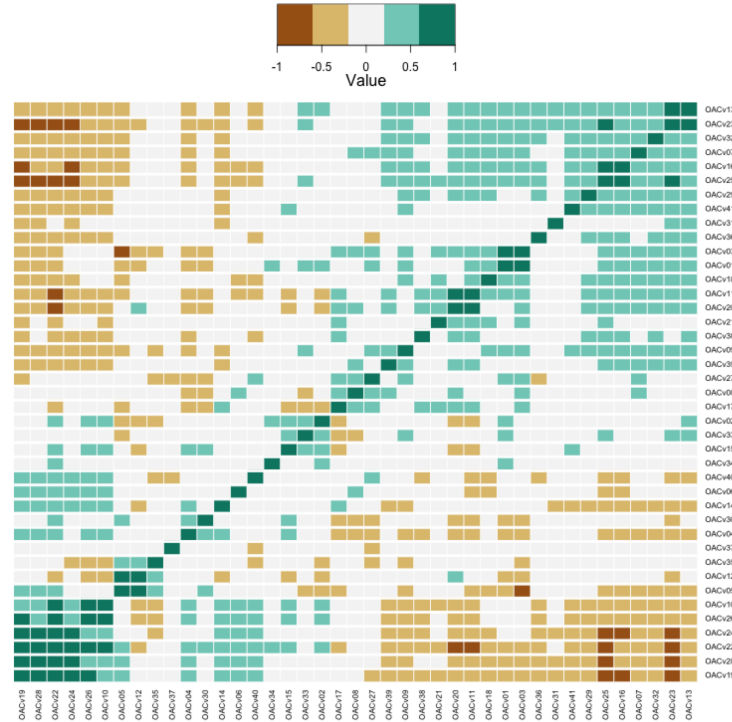


Figure 2.5: Compact Heatmap showing the correlation coefficient between all 41 variables of OAC 2001; using a test region of 500 OAs in four LA areas, ordered by degree of correlation.

scatterplot matrices are known as SPLOMs. Parallel Coordinates can also be used to present correlation between variables, although SPLOMs are typically easier to use for correlation (Munzner, 2014). Recent research by Harrison et al. (2014) investigates the perception of correlation in nine commonly used correlation visualisations and shows that scatterplots perform the best overall; however, parallel coordinates perform just as well for negative correlations.

Scatterplots can be encoded with additional information, for example they can display subsets of the data, such as grouping the data by geographical region, via the use of different symbols or colour (Reimann et al., 2008) or linked through interaction with other views (Monmonier, 1989). In high-dimensional datasets, dimension reduction is often necessary (Turkay et al., 2012) and SPLOMs are shown to be the safest idiom for representation (Sedlmair et al., 2013). Scagnostics, a term derived from the words scatterplot diagnostic, can be used to determine unusual structures in high-dimensional data (Tukey and Tukey, 1985; Tukey, 1993 in Wilkinson, 2005). For a large number of variables a compact heatmap (Munzner, 2014) is often used to represent the global correlation coefficients, as is the case in some examples of generating geodemographics, e.g. Fig. 2.5 was created using the geodemographic R package *gd* (Singleton, 2012) and similar matrices were used for variable selection in OAC 2001 and 2011 (Vickers, 2006; Gale, 2014b). As part of the statistical analysis, Gale (2014b) used a threshold of between

+/- 0.6 and +/- 0.7 to identify the highly correlating variable pairs, as there is no stated rule for a particular value indicating the switch between moderate and strong correlation coefficient.

Compact heatmaps for correlation with SPLOMs and histograms are used in DimStiller for variable dimension analysis and reduction (Ingram et al., 2010) and compact heatmaps, along with scatterplots, histograms and boxplots are used to visualise variable correlation and distribution in the feature selection visual tool by Seo and Shneiderman (2002; 2005).⁴ These are both good visualisation examples for aiding variable selection for geodemographic classification. However, the aspects of comparing both geographical variation and data scale are missing.

2.4.3 Geographical Variation

As geodemographics are inherently geographical, a variable with little or no geographical variation would have little influence on national clusters. To illustrate geographical distribution, choropleth maps are often used as the main form of visualisation as they quickly highlight if there is a geographical (or spatial) pattern in the data by using sequential scale of colour (Harrower and Brewer, 2003) to distinguish high from low values. The comparison of a pair of variables can be accomplished simply through the comparison of a pair of maps or by calculating the difference and mapping the values with a diverging colour scheme (Harrower and Brewer, 2003). Multivariate comparison can be achieved using multiple maps, although the number of variables is limited to screen space and scale can be a problem as map legends need to be comparable (Reimann et al., 2008).

The mapping of classes, using a quantitative scale (Harrower and Brewer, 2003), is common place in geodemographics, with examples dating back to Booth's London Poverty map from the 19th Century (Harris et al., 2005). There have been many examples of how maps illustrate variable relationship, together with amendments to the traditional choropleth style which tends to mask densely populated areas (as these are often the smallest areas on the map). Various alternatives have been designed to deal with this problem, in the form of population density-normalising cartograms (Tobler, 2004). Some preserve geographical shapes and space, like Vickers et al.'s (2010) Gastner Cartogram of OAC 2001 (see pp.410, Vickers et al. 2010), while others use rectangles (Wood and Dykes, 2008) or circles (Dorling, 1996). Slingsby et al. (2009; 2010b; 2010a; 2011) use rectangular cartograms arranged spatially, also termed spatial treemaps, to represent OAC 2001 Super Groups by colour and lightness to represent cluster uncertainty (Fig. 2.6). The use of rectangles instead of geographical regions allows for nesting hierarchical data and

⁴e.g. see Hierarchical Clustering Explorer, Fig. 1 in Seo and Shneiderman 2005, pp.98.

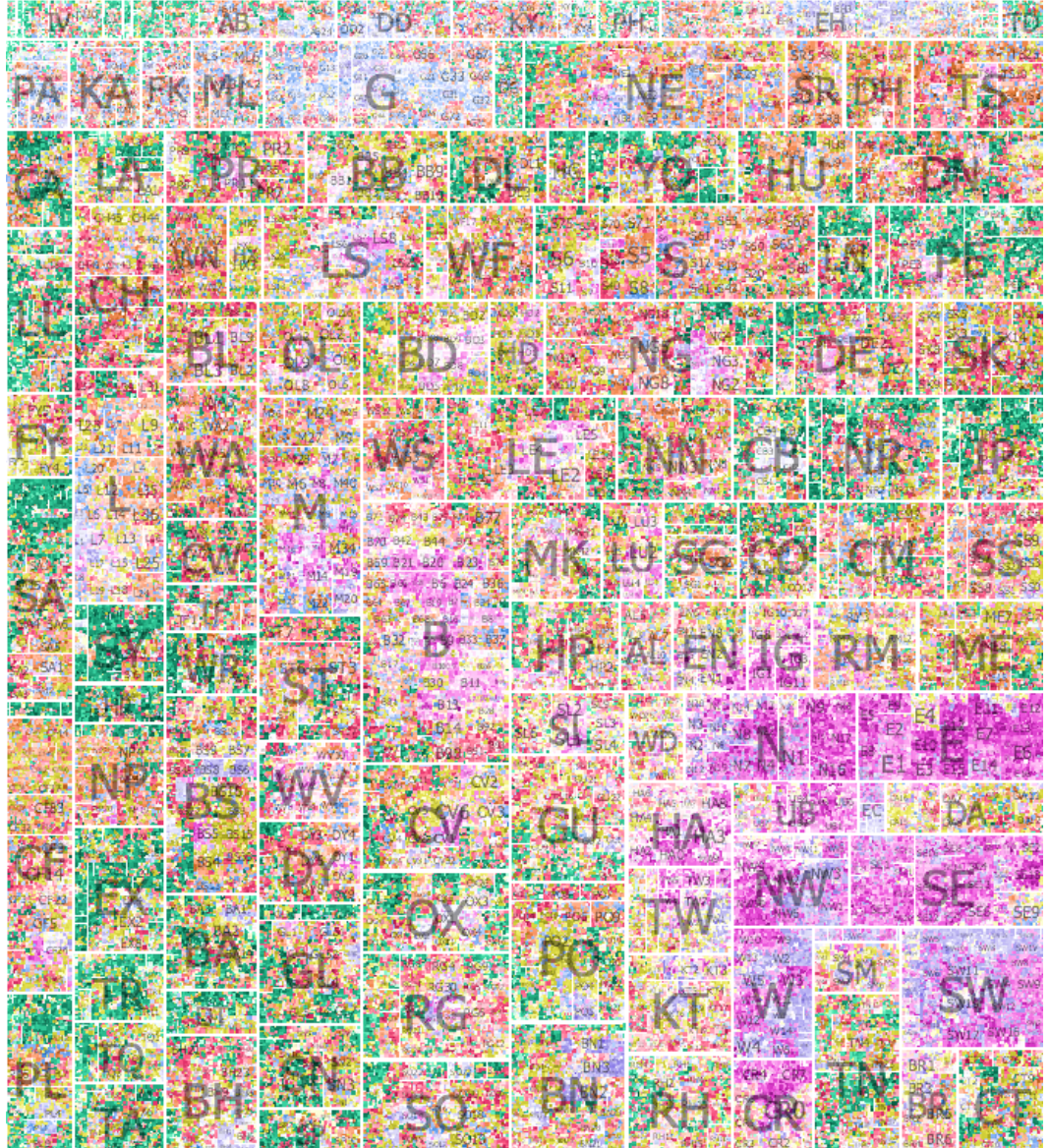


Figure 2.6: Hierarchical rectangular cartogram of all OAC 2001 Super Groups in all OAs, organised by postcode hierarchy. Hue denotes OAC super-group and lightness indicates uncertainty. Fig. 5 from Slingsby et al. (2011, pp.2549)

therefore allows for the inclusion of not only the super group classes, but also the groups and subgroups.

For OAC 2011 maps are used to visually assess the geographical distribution of 6, 7 and 8 clusters in order to determine the optimal dataset to use for the process, as shown in Fig. 2.7. In addition, a qualitative method was used to determine the geographical distribution of the individual variables. Twenty-five urban locations across the UK were chosen and percentages of each variable in each location investigated. While this is a useful overview, there are some drawbacks of this method. Not only is it biased to the chosen locations, but it does not distinguish between whether the percentage of the variable is concentrated in few OAs or dispersed across the whole urban area. This is

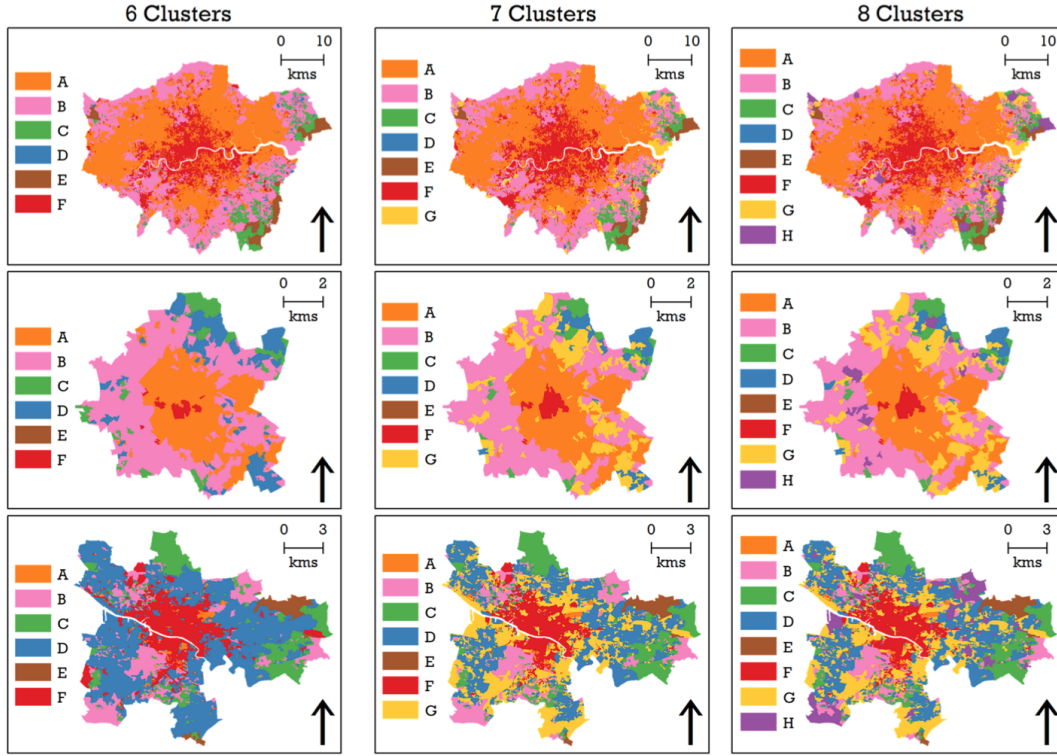


Figure 2.7: Classification maps of three urban areas: London (top), Wolverhampton (middle) and Glasgow (bottom) for one (Dataset 2) of four datasets compared for geographical distribution of clusters 6 to 8 cluster solutions for OAC 2011, Fig. 7.13 from Gale (2014b, pp.260), contains National Statistics data ©Crown Copyright and Database Right 2014, and OS Data ©Crown Copyright and Database Right 2014.

termed *spatial auto-correlation* – an indicator of how dispersed or clustered a variable is in geographical space (Anselin et al., 2002). Many of the OAC 2011 variables required individual investigation to interpret the geographical distribution, as small concentrated populations were seen as important to the characteristic of the area (Gale, 2014b).

While maps can be created for each variable and compared visually, there have been recent advancements in statistical methods for the investigation of geographical variation algorithmically. The spatial auto-correlation can be calculated in a number of ways. Moran’s I (Anselin et al., 2002) and Getis-Ord (Getis and Ord, 1992) are two examples which have both global and local alternatives. Global measures are useful for a summary of the spatial auto-correlation of the entire dataset, while local measures identify hotspots or local clusters within the dataset (Adnan et al., 2013). In terms of visual examples, Anselin et al. (2002) presents a linked example with multivariate maps, scatterplots and box-plots to represent local and global Moran’s I results. Global Moran’s I is also implemented as a function in common GIS software from ESRI (Anselin et al., 2002). Other geographically informed statistics research includes geographically weighted regression (Fotheringham et al., 2002) and geographic data mining (Miller and Han, 2009).

In terms of geodemographics, Singleton and Longley (2009a, pp.293) argue that the generation process is “*fundamentally at odds with a variety of developments in quantitative geography and computing*”. Such movements in geographically informed statistics are not included in the process and local ‘place effects’ are ignored (Singleton and Longley, 2009a). While the use of geographically weighted statistics and auto-spatial correlation has started to be investigated for geodemographics (Adnan et al., 2013) more research is needed. For analysing the geographical variation of the variables in the comparison process, the inclusion of geographically weighted summary statistics (Fotheringham et al., 2002; Harris et al., 2011) could enable the creator to better understand the geographical relationship of pairs of variables. There are growing examples of geographically weighted statistic visualisation including linked views of small multiple maps, with boxplots, scalograms or parallel plots (Dykes and Brunsdon, 2007; Harris et al., 2014).

2.4.4 Variable Scale

As mentioned in Section 2.3.3.2, data scale is of great importance when creating and deciding on variables for geodemographics. Classifications can be created at multiple scales, with each likely to produce very different outcomes. The nature of the summaries used to describe areas at each scale resolution, and relationships between them, can vary. It is known that geographical phenomena can vary with location and scale (Andrienko et al., 2010; Brunsdon et al., 1996; Fotheringham et al., 2002), and scale therefore needs to be a core component of multivariate geographical analysis. Lam and Quattrochi (Lamand and Quattrochi, 1992, pp.88) suggest opportunities for visualization when noting that analysis that does not account for scale may well be invalid and that “*analysing geographical phenomena using a range of scales offers a special view*”. The use of visualisation to illustrate how different variables react to changes in resolution can reveal patterns in the data (as shown in Turkay et al., 2014) and may help to identify the optimal resolution for analysis as well as illustrate the effects of MAUP.

2.4.5 Visual Parameter Space Analysis (vPSA)

The investigation and visualisation of multiple scales and the inclusion of geography in multivariate comparison is much more complex than standard multivariate statistical comparison, as there are multiple dimensions of the data to be compared. This type of analysis falls into the research area of *visual parameter space analysis* (vPSA), recently described by Sedlmair et al. (2014) as interactive visualisation that facilitates parameter space analysis (PSA), where PSA is defined as “*the systematic variation of model input parameters, generating outputs for each combination of parameters, and investigating the relation between parameter settings and corresponding outputs*” (Sedlmair et al., 2014, pp.2162). vPSA offers an opportunity to deal with the complexity, as well as keep the

human ‘in the loop’. This is relevant to a number of elements of the process of generating a geodemographic classification, in particular the variable selection process which is a time-consuming and manual process. Sedlmair et al. (2014) demonstrate that this is a growing area of research in the visualisation domain where 21 core papers, from an initial set of 112, were analysed and compared to create a conceptual framework which helps to abstractly describe vPSA problems across application domains. vPSA has three components: data flow model, navigation strategies and analysis tasks. There are four typical types of vPSA navigation strategies: *informed trial and error*, *local-to-global*, *global-to-local* or *steering*. Each are beneficial for different situations and sometimes used in combination. Six typical analysis tasks associated with vPSA are also identified from the 21 case studies: *optimisation*, *partitioning*, *fitting*, *outliers*, *uncertainty* and *sensitivity*.

Many of the vPSA analysis tasks can be associated with the process of generating a geodemographic classification, described in Section 2.3. Optimisation is relevant when investigating the clustering algorithm as well as the optimal standardisation, transformation/normalisation and preparation methodologies, as carried out statistically by Gale (2014b). Outliers in variables are important to identify as they may influence the clusters. These outliers can be removed through a method of transformation, implementation of a threshold or introduction of variable weighting. Much uncertainty is evident in the generation of geodemographic classifications and a visual approach to the process is likely to reduce uncertainty and improve understanding of the current ‘black-box’ situation associated with the clustering algorithms, data transformations and variable selection process. Investigating sensitivity is referred to as “*What ranges/variations of outputs to expect with changes of input?*” (Sedlmair et al., 2014, pp.2166). This is highly relevant to the investigation of the effects of scale (resolution or extent) or calculation of local geographical statistics on the variable inputs for the variable selection process and links the research for RQ3 and RQ4 to Sedlmair et al.’s vPSA model. Incorporating specifically designed geographical visualisation as a key component of vPSA can be referred to as *geo-visual PSA*, or shortened to *gvPSA*. The need for *gvPSA* is demonstrated and explored in this thesis, particularly in the exploration of the scenarios in Chapters 7 and 8.

2.4.6 Multivariate Comparison Summary

This section identifies many ways of calculating statistics and creating visualisations for the comparison of multiple variables. Multivariate comparison is described as much more complex than uni- or bi-variate comparison as there are many variables and data items to consider and view at one time. This act of comparison is one of the fundamental

tasks of many visualisation tools and there are many examples of best use. A visual interactive approach to aid the creation and in particular the variable selection process is missing in the geodemographic process and is proposed for this research. This visual approach will allow skewed and strongly correlating variables to be quickly identified and investigated and importantly for geodemographics, the geography of multi-scale correlation to be explored. Despite the lack of visualisation research for the variable selection process for the generation of geodemographics, the visualisation of multivariate data in general has been researched in detail. Multiple linked visual views such as maps, histograms, box plots and scatterplots together with the use of globally or locally weighted statistics would allow the user to gain more information, enabling transparency of the variable differences and the process itself. The visual comparison of multivariate data across scale and with the inclusion of geography forms a complex parameter space and can be linked to vPSA, where the design of visualisation and mapping of visual variables is important.

2.5 Chapter Summary

This review of a broad range of literature from various disciplines reveals that the energy industry is rapidly changing, and data visualisation and visual analytics are proving beneficial to both energy consumers and analysts within the industry. There is growing demand for more advanced data analysis with the increased pressure for energy reduction and better efficiency. The profiling of energy consumers by typical traits will allow for more targeted marketing, specialised services and specifically tailored tariffs. Technical advances in the industry create a path for more advanced profiling based on smart grid or smart home data.

Certain variables are recognised in the research to relate to energy use, in particular disposable income, housing type, tenure and rural/urban location. As the consideration of geography and demographics is important, an energy-based geodemographic classification is proposed. The generation process is, however, shown to be particularly labour and time intensive, where expertise in the methods and processes are advantageous. While some aspects of the process of generating a geodemographic classification have been visualised, there are aspects, in particular the variable selection process, which would benefit from more advanced research and specifically designed visualisations to aid the process. The process relies heavily on multivariate statistics and there is a lot of relevant research in this area.

The inclusion of local statistics to investigate geography introduces a complex parameter space where more research on visual possibilities is needed. A number of data uncertainties and sensitivities related to data scale are also identified in the process, which could be displayed and highlighted in visualisation through the use of visual

variables. In summary, specifically designed data visualisation to aid the understanding of the variable selection process for the generation of geodemographics and allow for the consideration of geographical variation is an area where research and examples are lacking. Core to this activity is an investigation into the varying effects of data scale and geography in multivariate comparison along with access to techniques that support this process. Creative approaches to data use and visual design are shown to be useful, and are likely to be beneficial to the design process.

3

Creative Energy Domain Exploration

The research presented in this chapter describes a creative requirements gathering and evaluation process with energy analysts. The research investigates future and current opportunities for data analysis and visualisation solutions for energy companies. This links to the first two motivational questions described in Section 1.2: “*What is the future for household energy analysis?*” and “*What value can be derived from energy consumption data through data analysis and visualisation?*”. The research includes the use of creativity techniques to deliberately stimulate creative thinking within the Requirements Workshop. The use of these techniques was evaluated within data visualisation methodology and the findings were published in IEEE TVCG (Goodwin et al., 2013) and presented at the international IEEE InfoVis 2013 conference (see Appendix A.1). The results demonstrate that there is a need for data visualisation within energy companies as well as the need for creative, novel, yet comprehensive data visualisation designs. The need to analyse energy users and create profiles based on typical traits is also seen as important for improving understanding of consumer behaviour. The clear need for profiling energy consumers prompts continued research in accordance with the third motivational question: “*Is there a need for an energy-based geodemographic classification?*”, which is investigated in subsequent chapters.

3.1 Smart Home Project Description

The research described in this chapter forms part of a joint initiative between City University London and the IMDEA Energy Institute, Madrid. The research project titled ‘*Visualizing the Smart Home: creative engagement with customer data*’, referred to as the ‘Smart Home Project’ in this thesis, was funded by the 2012 E.ON AG International Research Initiative (IRI) (E.ON Technology & Innovation, 2014). This project was undertaken in collaboration with energy analysts (Forward Thinking Technologies Team, E.ON), energy data modellers (City University London’s Electrical and Electronic Engineering Department and IMDEA), visualisation designers (giCentre, City University London) and creativity experts (Centre for Creativity in Professional Practice, City University). The overall project involved four strands of work:

1. building a smart home model;
2. discovering data analysis and visualisation opportunities for energy companies;
3. optimising the smart home model;
4. discovering data analysis and visualisation opportunities for the customers.

Together the results of these four strands were combined to discover creative ideas for new services for the energy industry based on using and visualising smart home data.

The author of this thesis managed the work related to the second strand of work where smart home visualisation possibilities for the energy companies were investigated. This involved joint-running the requirement and design workshops, and the feedback and evaluation sessions as well as documenting the analysis, results and process. Activities included judging the feasibility of the requirements, collaborating with the developers, interviewing the modellers, transcribing the feedback and evaluation sessions and analysing the results. The techniques for the ‘Requirements Workshop’ and the evaluation of the creativity were chosen with advice from experts in the area¹, and the visualisation prototypes (described in Sections 3.5.2) were developed by other members of the giCentre². The analysis and results of this research form the initial stage of research for this thesis (see stages of research in Fig. 1.1).

3.2 Creative Requirements Workshop

The aim of the workshop was to gain an overall insight into the analytical and visualisation opportunities which are arising with technological advances and increased data collection

¹Creativity Experts: Alison Duffy, Sara Jones, Graham Dove and Amanda Brown from Centre for Creativity in Professional Practice, City University london

²The prototypes were developed by: Aidan Slingsby, Alex Kachkaev, Jo Wood and Iain Dillingham

in the energy industry. The use of creativity techniques was deemed appropriate as smart home technologies are still being developed, so the possibilities for analysis are vast and have broad potential. In order to keep requirements open and to invite participants to have new and potentially novel ideas, time was taken to establish a creative atmosphere and specifically chosen techniques were used to stimulate creative thinking. Working with creativity experts (City University London), techniques from methodologies such as creative problem solving (CPS) (Osborn, 1957) and Synectics (Gordon, 1960) were considered for use in the Requirements Workshop and contemporary innovation literature was consulted (Hohmann, 2007; Michalko, 2010).

Two pilot sessions were run internally at City University London, where creativity techniques were tested and adapted with members of the project team. The ‘I wish’ activity (McFadzean, 1998) was found to be too abstract given the need to concentrate on data analysis and visualisation possibilities and was adapted to three separate prompts; “What would you like to *know*?”, “What would you like to *be able to do*?” and “What would you like to *see*?”. An option for the use of a metaphor (Duffy, 2013) in the shape of a tree to place statements on, to allow participants to identify the ‘low-hanging fruit’, was found to be too restrictive to the exercise as options needed to be kept open and flowing. The prioritisation of the ideas was only informally approached during the final storyboarding activity. Feasibility and prioritisation of the ideas was instead left to the researchers and designers with experience of the data and data visualisation possibilities.

As well as tailoring the creativity techniques, careful attention was paid to the choice of venue, as the physical environment in which activities are carried out can have a significant impact on the creative climate (Isaksen et al., 2011). The day-long workshop was held in September 2012, at a neutral venue, with plenty of space, light and refreshments and was attended by five participants from the Forward Thinking Technologies team at E.ON UK. These five smart home experts (referred to throughout as energy analysts) work together on a regular basis and are often involved in thinking of new ideas and possibilities for smart home technologies. It is therefore possible that emphasising creativity was not required as much as in other cases studies (e.g. Pennell and Maiden, 2003; Maiden et al., 2007); however, the analysts’ knowledge of the new datasets available and the opportunities offered by data visualisation were limited.

The day began with some warm up activities, including an introduction that encouraged trust-building and participation, while introducing some analogical thinking. Participants were asked “*if you were to describe yourself as an animal, what would you be?*”. This encouraged individual thinking and some interesting responses ranging from a Meerkat “*head up, eyes open, ready to take everything in*”, a Buzzard “*overlooking everything and spotting what things to swoop down on*” and a Squirrel “*keeping busy by*

3.2. CREATIVE REQUIREMENTS WORKSHOP

Activity	Aspiration Topic	Established	Feasible
Know	Customers Habits	10	5
Know	Appliance Consumption	6	6
Know	The Value of the Data	2	2
Know	Visualisation Design	2	2
Do	Improve Customer Experience	5	2
Do	Manage Energy Demand	3	3
Do	Advance the Technology	3	0
See	Data Analysis & Visualisation	8	6
See	New Products and Services	1	1
—	—	-	-
Next?	Change Customer Behaviour & Improve Life	5	0
Next?	Improve & Expand the Product	6	0
Next?	Understand Customer Habits	3	2
Next?	Gain Trust & Increase Customers	5	0
Next?	Educate Energy Industry & Manage Demand	5	1

Table 3.1: Aspirations revealed in ‘*Know/Do/See*’ and ‘*What next?*’. Numbers show total aspirations established and those deemed feasible. Table 1 from Goodwin et al. (2013, pp.2519)

gathering everything together. Statements and quotations emphasising creativity and thinking with new eyes were also shared such as Einstein’s famous quote: “*If at first, the idea is not absurd, then there is no hope for it*”. Four activities were run during the day by a professional facilitator, with each activity adding to the ideas of the previous: Wishful Thinking, Constraint Removal, Visualisation Awareness using Analogical Reasoning and Storyboarding. Photos of the workshop activities are shown in Appendix B.2 and each is explained in greater detail in the following sections.

3.2.1 Wishful Thinking

To capture aspirations or what Hohmann (2007) describes as ‘opportunity statements’ the analysts were asked to think about the data becoming available in the wider context of the smart home programme: “What would you like to *know?*”, “What would you like to *be able to do?*” and “What would you like to *see?*” Participant’s aspirations were identified individually in a brainstorming (Osborn, 1957) exercise, recorded on coloured post-its relating to each theme (Know, Do and See) and then explained to and discussed with the group. Post-its were placed on flip-charts and similar or overlapping items were grouped together. Part two of the activity pushed participants further in their thinking by asking them (in small groups) to consider an aspiration and think about what would happen *next?*, assuming that the chosen aspiration had already been achieved. Subsequent ideas were also recorded on post-its and placed on flip-charts which were hung on the wall, in order to be seen for the next (and subsequent) exercise. In total, 64 aspirations were established and were later grouped into similar topics (see established and aspiration topic in Table 3.1). On reflection it was noted by the research team that the items in ‘See’ and ‘Know’ reflected ideas which could help inform the visualisation design, while those in ‘Do’ gave a contextual background, which could aid the evaluation of the designs.

3.2.2 Constraint Removal

The second activity had both a divergence and convergence component and was based on previous work by Jones et al. (2008). The activity began by everyone calling out the current barriers or constraints to achieving the aspirations identified in the first exercise. These were noted on a flip chart and when the list of constraints was exhausted, a *constraint* was taken in turn by small groups of participants and they were asked to identify what new possibilities would arise if this barrier/constraint no longer existed. Again ideas were generated and recorded.

Constraints identified included: technical and hardware issues, resource and time limitations, customers being difficult to understand as people lead complicated lives, a lack of customer trust, and conflicting business priorities. Removing these constraints unlocked a number of ideas for moving forward. In particular, deriving value and knowledge from the smart home data, allowing the improvement of services, and thus gaining both customer and industry trust. These constraints and additional aspirations did not produce any new design or evaluation requirements for the project team; however, they were particularly useful for building context and understanding. This not only highlighted the potential to use visualisation to engage with the customers but also to engage with the business stakeholders.

3.2.3 Visualisation Awareness using Analogical Reasoning

The technique of ‘Analogical Reasoning’ that can help develop novel or unexpected ideas and assist participants in refining ideas through ‘mental stretching’ (Isaksen et al., 1999), has been shown to be useful in creativity workshops (Maiden et al., 2007). This technique was combined with a ‘Visualisation Awareness’ activity which has been found to be beneficial in data visualisation requirements gathering workshops (Dykes et al., 2010; Koh et al., 2011). Prior to this activity a lunchtime ‘Excursion’ (Gordon, 1960) or ‘*Imagery Trek*’ (Osborn, 1957) introduced the concept of analogical reasoning, where participants were asked to bring back an object which had a link to the smart home programme. This maintained creative energy and outlooks over the break as well as enabled participants to contribute, communicate and experiment with analogy. The brief exercise had some surprisingly creative responses. One participant brought back a copy of ‘Great Expectations’ found in the lunchroom, another took a photo of the surrounding farmland and described it as *“not natural at all but it has had many years of man-management”*, whilst another participant described the feeling of the sun: *“it is nice and warm, comfortable and secure, but it is slippery and difficult to get the temperature right”*.

During the visualisation awareness activity a slide-show of different examples of visualisation techniques from various domains were shown to participants. This included

demos of visual and analytical software solutions, videos of data storytelling as well as static or animated visual imagery. Participants were asked to think of connections that related each to the energy industry. Reactions were again recorded and placed on the wall. In total ten analogical ideas were inspired by the visual demos. This included an idea for using bubbles of consumption increasing and decreasing when used in the home, influenced by *Empires Decline Revisited* (Cruz, 2010) and an idea to show wasted energy flows sparked by an animated visualisation of millions of bike journeys (Wood, 2012).

The bubbles of energy consumption were particularly interesting to participants as their trial data was seen as relatively uncertain. What was described as the “*wobbly factor*” of the bubbles in *Empires Decline Revisited* inspired thinking about how to represent data uncertainty. While the accuracy of the data had already been mentioned as very important to the participants, with “*I would like to see pretty, but precise data*” being an aspiration from the wishful thinking activity, the visual activity prompted new ideas on the importance of visual design: “*The design is as important as the information*”. Design-wise, it was also seen as important that visualisations were *beautiful, engaging* and *simple*, for instance to be able to do “*everything in three clicks*”.

Data analysis opportunities were discussed with a notable demand to group, filter and compare data by appliance type, temperature, user demographics, time and geography to better understand consumption variability across types of users and different uses. The session ended with a highly creative and motivational *Plan of Action* for the focus of smart home data analysis:

1. *Discover*: find out where energy is used
2. *Displace Consumption*: change behaviour and control devices
3. *Reduce Energy Production*: specifically by the amount needed to close a power station³

3.2.4 Storyboarding

In this session the aspirations, constraints, analogical ideas and design requirements from the three previous activities were brought together. Blank storyboards of varying types were provided, with images of the visuals, which were shown in the Visualisation Awareness demo (prepared prior to the workshop for cutting out and using for storyboards, when appropriate) along with coloured pens for sketching. Storyboarding is often used in both visualisation and creativity workshops to bring ideas together (Maiden et al., 2007, 2004). Here, ideas from the day were sketched onto storyboard templates using the title; “*A day in*

³A power station is an industrial facility for the generation of electric power. Also referred to as a power plant, generating station or generating plant

the life of an E.ON analyst". This allowed participants to informally prioritise aspirations and highlight visualisation ideas which they could identify as being particularly beneficial to their daily work.

Due to the previous exercise being particularly creative, there was limited time left for the storyboarding activity but stories were produced and presented to the group relatively quickly. One of the analysts chose to create a mind-map of ideas, which was also useful for us to see informal prioritisation. The storyboards and mind-map are shown in Appendix B.3. In general, key themes that were evident from the storyboards emphasised the need for greater understanding of consumer habits and the desire to understand customer behaviour by grouping and comparing relevant data.

3.2.5 Requirements Workshop Outcomes

In general, the outcome of the workshop's activities identified five main themes which were seen as important to the continuation of the smart home programme:

1. *Analyse the Data* to understand more about customers' energy habits and appliance consumption;
2. *Develop Knowledge* to start to prove/disprove myths and theories of energy saving and behaviours;
3. *Communicate and Engage* within the business, and with industry and the general public to manage demand and change behaviours;
4. *Build Trust* in the company and the products;
5. *Improve and Expand Smart Products* beyond energy to improving comfort and security.

One of the initial aspirations; "*I would like to know the who, what, when, where and why of energy consumption*" sums up the first of the five themes. While *why?* is difficult to determine from quantitative data, the *who, what, when and where?* are typical characteristics to investigate for analysis. There is evidently a clear need from the industry for further research to analyse data and discover as much as possible about energy use. Using the requirements gathered from the workshop, a concept map was produced to map connections between data associated with the energy industry in light of data available for the project (see Appendix B.4). The concept map reveals that many aspects of energy analysis link directly to consumer profiling, demographics and lifestyle, yet much of this is unavailable for the smart home project due to limitations of the datasets (as explained in the following section).

Following the workshop, the aspirations, barriers, analogical ideas and storyboards were collected and assessed for feasibility for smart home visualisation given the data available (see Section 3.3) and in preparation for further work. From the 64 aspirations from the Wishful Thinking activity, only 30 were identified as being potentially feasible and therefore important to highlight during the next stage of the process. The final column in Table 3.1 identifies the number of feasible aspirations in relation to the topics. More detail is also available in Appendix B.3.

3.3 Smart Home Data Availability

Two sources of smart home data were available for the research: live data from a smart home trial and modelled data simulating future scenarios.

The live data contained electricity and gas consumption for all appliances of a test-bed of 130 properties participating in a smart home trial, including 75 with metadata relating to resident demographic variables and geographical location. The trial dataset contains data from smart meters and smart plugs, totalling more than 18 million records over a period of 14 months. The data has particularly challenging characteristics: irregular timings, variance of frequency of recordings from minutes to days, and differences in the number and selection of appliances. The sample was also biased in terms of demographics and geographical location as it was a small self-selecting sample group.

The second dataset, developed as part of the (EON IRI funded) smart home project (Gruber and Prodanovic, 2012), generates appliance-based energy usage scenarios at 15-minute intervals for any given period for a given number of households. This represents large scale scenarios of what data from smart homes could be like in the future. The values in the model are based on real survey data available in a detailed UK government report (Energy Saving Trust, 2012). The results are limited to weekend and weekday activity, but include daily and seasonal variations of consumption, standby options and optimisation calculations to simulate shifting and reduction of demand over time. A drawback of the model (in respect to the analysts' aspirations), is that appliance use and distribution of appliances to households are determined probabilistically, therefore when analysing individual households there is an unrealistic appliance-ownership relationship and typical usage pattern. The simulation also does not include any household demographics or geographical location.

Both sources contain numeric information for total electricity and gas consumption as well as individual appliances for each household (real or modelled) along with time of recording. Derived values (average, max, min, count, standard deviation) were calculated in both cases by groups of time period (hour, day, week, month) and by category (such as appliance type). Despite the aspiration for needing to discover the *who*, *what*, *when* and

where? of energy consumption, the data is limited to discovering the *what?* and *when?* of energy consumption, as the *who?* and *where?* are limited to a small and biased sample in this instance.

3.4 Design Workshop, Development & Feedback

Following the requirements workshop, a ‘Design Workshop’ was held by the author and Prof. Jason Dykes, with five other visualisation specialists from the giCentre, in order to gain ideas and designs for building visualisation prototypes appropriate for the future of smart home analysis. The background of the project first was explained, feasible requirements, sketches and ideas from the requirements workshop introduced and the content of the available datasets outlined. Working in pairs, design ideas were linked to the aspirations. They were then investigated, sketched and discussed. More details and photos are available in Appendix B.5.

From the design ideas, four prototype visualisations were designed and developed by four giCentre members, in parallel over a one month period, with two iterations of rapid agile development. Rapid and parallel prototyping (Dow et al., 2010) was seen as appropriate for this project as there were many possible outcomes from the requirements workshop and restricting ideas to one solution would not enable the analysts to see the vast possibilities and benefits of data visualisation – an ultimate goal of the project. The first iteration of development commenced after the Design Workshop and prior to analyst feedback, while the second followed a ‘User Feedback Session’ and after prioritising the feedback.

During the User Feedback Session, each prototype was explained in turn by initially identifying the relevant aspirations and ideas from the Requirements Workshop and subsequent designer/developer questions from the Design Workshop. Each prototype was then demonstrated and initially driven by the author before control was passed to the users. The feedback session was audio-recorded and comments and enhancements were noted. Possible enhancement opportunities identified (and development time previously estimated) by the project team were explained (see Presented in Table 3.2) and prioritised with the analysts using a scale of 1 to 10, where 10 had the highest need. Further user-suggested enhancements (see User-Suggested in Table 3.2) were also prioritised. Enhancement feasibility was derived from combining the analysts’ prioritisation (business need) with the developers’ estimates for difficulty of implementation (Cohn, 2005). During the iterations, features were also prioritised using the MoSCoW technique (Brennan, 2009). Several enhancements were implemented for each prototype (see Implemented in Table 3.2). Other agile development techniques were also followed

3.5. EVALUATION: PROTOTYPE DESIGNS, CREATIVITY & PROCESS

Prototype Name	Presented	User-Suggested	Implemented
<i>Demand Horizons</i>	11	7	6
<i>Consumption Signatures</i>	7	5	10
<i>Ownership Groups</i>	10	3	8
<i>Smart Home HeatLines</i>	10	3	6

Table 3.2: Prototype Enhancements, Table 2 from Goodwin et al. (2013)

Considering	Evaluating	Method
The Prototypes	Appropriateness	Questionnaire
The Prototypes	Novelty	Structured Group Discussion
The Prototypes	Surprise	Structured Group Discussion
The Design Process	Validity & Effect	Structured Group Discussion
The Design Process	Creativity	Reflection by Designers

Table 3.3: Evaluation Process, Table 3 from Goodwin et al. (2013)

for the process, such as daily meetings between the author and the designers/developers to re-prioritise and discuss design decisions in light of requirements and barriers.

The final four prototypes, named ‘Smart Home HeatLines’, ‘Consumption Signatures’, ‘Demand Horizons’ and ‘Ownership Groups’ are explained briefly in the following section, with further details available in the TVCG paper (Goodwin et al., 2013, – in Appendix A.1). User interaction of the prototypes is also shown in the accompanying video ⁴.

3.5 Evaluation: Prototype Designs, Creativity & Process

A structured process was constructed to determine the extent to which both the prototypes and design process were seen as valid and creative. Table 3.3 divides the evaluation into five processes. The evaluation processes used a combination of visualisation evaluation techniques (Sedlmair et al., 2012) and creativity evaluation techniques (Maher and Fisher, 2012; Dean et al., 2006). Details are discussed in the following section.

3.5.1 User Evaluation Session

The ‘User Evaluation Session’, with four of the five energy analysts who participated in the original Requirements Workshop, was designed to evaluate the four prototype designs for *appropriateness*, *novelty* and *surprise* (see Table 3.3), as these have been identified as most commonly used for the evaluation of creativity (Dean et al., 2006). The session began with a presentation of each prototype, while demonstrating the new enhanced functionality through (increasingly analyst-directed) chauffeuring. The features were again linked to the original requirements and the feedback. Chauffeuring, rather than passing the control to the users, was deemed appropriate as a means of getting analysts to use the software in a timely manner and because the project was not evaluating the usability of the prototypes but the value of the approaches themselves.

⁴InfoVis supplementary video featuring all four prototypes available in the digital appendix and online at: <https://vimeo.com/69185134>

After each demonstration a questionnaire (available in Appendix B.6) was completed by each of the analysts, evaluating the *appropriateness* of the prototypes by assessing the extent to which the prototype satisfied the relevant requirements, on a scale ranging from strongly agree (1) to strongly disagree (6). These are shown in Fig. 3.1 and explained in the sections to follow.

Novelty and *Surprise* were then determined from a structured group discussion. During the discussion, the prototypes were again used through direct chauffeuring to prompt discussion on novelty of the design, potential surprise they encountered and particular data insights. Discussion prompts included asking the analysts to think about who would find the tools useful and for what purposes and whether there were any problems, opportunities or improvements to be made. Questions regarding the design process itself were also prompted. The transcript from the User Feedback Session was used together with the transcript from the Evaluation Session to investigate novelty and surprise, as surprise tends to wear off on the second viewing. To evaluate this, two researchers with knowledge of the project separately coded 109 quotes from the workshops with the terms: ‘Novelty/Surprise’, ‘Appropriateness(of design)/Value added’ and ‘Data Insight’. Examples include; Novelty/Surprise: *“I like the amount of interaction ... there are just so many touch points ... it gives you a wow factor!”* and *“It’s 18 million data points!”*. Appropriateness/Value: *“[using this] we could start to target things differently”* and *“This would be invaluable in starting to prove that some of these electronic approaches work”*. Data Insight: *“This [dryers had a later peak then washing machines] goes against the insights that people don’t like to leave the washing in”* and *“This chart measures the consistency of our data collection [and shows where there are gaps]”*.

Novelty and surprise were merged into one category as it was found to be too difficult to determine the difference. Quotes placed in this category referred to comments referring to surprising or novel aspects of the visual design. When quotes were clearly referring to surprising data insights these were coded as ‘Data Insight’ rather than ‘Surprise’. As insight is one of the primary functions of visualisation and analysis tools (Saraiya et al., 2005) it was also seen as important to keep these separate for our evaluation. There is a fuzzy overlap between the codes used, which meant that further analysis and evaluation of the structured discussion was difficult. Yet, all four prototypes revealed some aspects of novelty and surprise in the design, with three of the four being deemed appropriate to the requirements. Each prototype is described in turn in the following section.

3.5.2 The Four Prototypes

The design, functionality, feedback and evaluation of each of the four prototypes is briefly explained in the following sections.

3.5. EVALUATION: PROTOTYPE DESIGNS, CREATIVITY & PROCESS

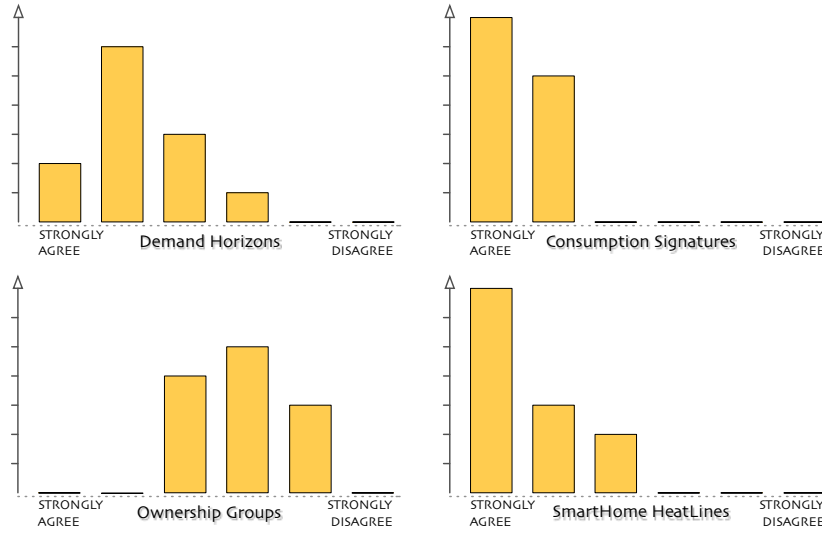


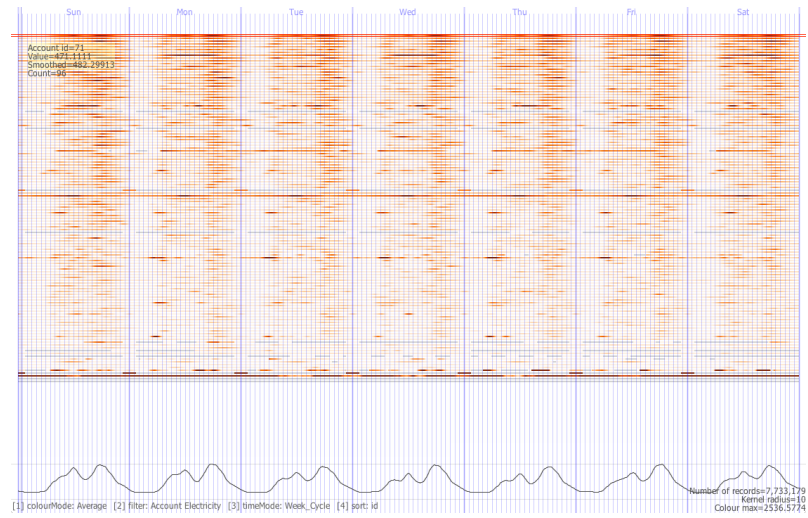
Figure 3.1: The responses to the prototype ‘appropriateness’ questionnaire, ranging from strong agreement (1) to the left and strong disagreement (6) to the right – Fig. 5. from Goodwin et al. (2013, pp.2521)

3.5.2.1 Smart Home HeatLines

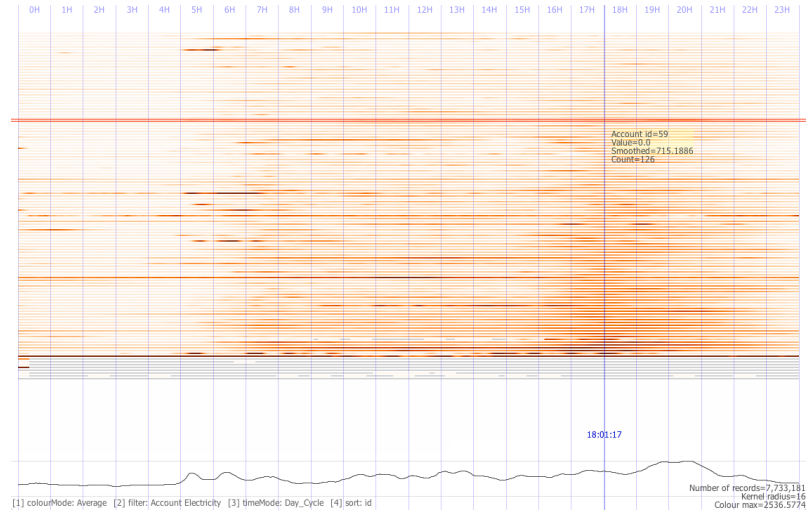
This was the only prototype which presented E.ON’s own trial data and this was a huge positive in terms of excitement and engagement from participants as real patterns and anomalies were identified through visualisation during the tool demonstrations. The design allows the data for each household (vertically) across time (horizontally) to be explored and re-aggregated on the fly, allowing the large volumes of data to be displayed. Fig. 3.2 demonstrates some of the possible views of the prototype which uses over 7.5 million data points.

Additional features to group the households by type (whether house type, number of rooms, number of residents or heating type as in Fig. 3.2c) were specifically requested during the feedback stage and these were added during the second development stage. This was found to be extremely useful to give context to the visual findings: *“it gives us a whole new way of analysing people”*. A map (not shown in Fig. 3.2 for data privacy) to show the data geographically was also seen as a useful additional view as it helped to give known context to the otherwise quite abstract visual representation. Clustering profiles of households based on the similarity of their consumption profile over time (e.g. Fig. 3.2a) was also a second iteration enhancement, chosen by the design team, and was seen as very useful by the participants, who had not expected to see so many diverse consumption patterns between the households.

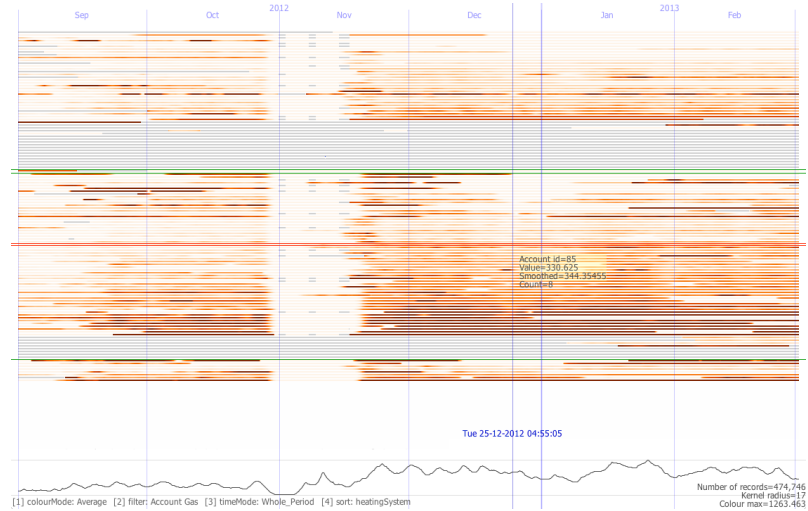
All appropriateness scores were between 1 and 3 indicating that the prototype was considered highly relevant to the analysts’ needs. The tool was seen as appropriate for *“a very wide user base”*, in fact *“anyone interested in gaining insight from energy consumption*



(a) Average weekly electricity consumption ordered by similarity to selected profile



(b) Average daily (24 hour) electricity consumption ordered by max use at 6pm



(c) Average monthly (Sep-Feb) gas consumption ordered by heating type (green bars) and max use on 25th Dec

Figure 3.2: Screenshots from Smart Home HeatLines showing energy consumption for each (trial) household (vertically) over time (horizontally)

data". The tool prompted many insights relating to individual participants as well as general patterns of energy use and also data anomalies, such as the clear band of missing gas data for the first two weeks of November (as shown in Fig. 3.2c). In general the tool was seen as beneficial as it would greatly improve communication of the smart home project amongst colleagues. The value of exposing the analysts to the trial data in this way was explicit: *"this would be invaluable in starting to prove that some of these electronic [smart home technology] approaches work."* This creativity-informed data prototype not only got the analysts excited by their data but changed the analysts' expectations, needs and ambition for continued work.

3.5.2.2 Consumption Signatures

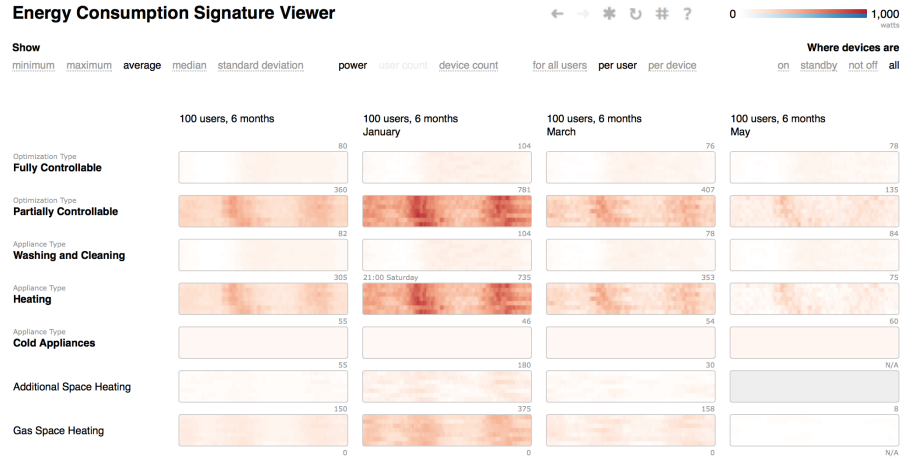
This prototype used a familiar technique of heat-mapping in a new context. The slick and smooth representation of large amounts of data simply through the use of colour variation and column/row reordering to change the comparison and view was seen as very impressive and *"very clever"*. This prototype was seen to have the most potential for use across the whole industry in terms of educating people on the variety between the different appliance consumption patterns over time. The data was grouped into different types: appliance, appliance type, detailed appliance type, room (where the appliance is usually used (e.g. Kitchen), greenwave class and sub class (as defined by E.ON), energy pattern type (see Fig. 3.3a) and optimisation type (see Fig. 3.3c) as created and defined by the modellers (Gruber and Prodanovic, 2012). Clear patterns can be identified by time (by hour, week, month and season – see Figs. 3.3a-3.3c), optimisation potential (see Fig. 3.3c) as well as number of households in the comparison datasets (see user counts in Figs. 3.3a-3.3c). Consumption patterns are shown through derived statistics (mean, medium, standard deviation, maximum, minimum) and by distinguishing power from users as well as devices. Standby is also identifiable and patterns can be identified when standby consumption is separated from total consumption and when it is switched off – a feature which excited the analysts as standby is a known issue in the industry for wasted energy.

The tool scored 1s and 2s in the questionnaire, with a mode value of 1 signifying very strong agreement that requirements were satisfied. It was seen as *"very powerful and very useful"* and a knowledge building tool: *"I could imagine ... just taking a week off and just letting your curiosity dive in and out."* Expected (and unexpected – see Section 3.5.3) patterns were easily identifiable which increased the participants trust in the modelled data and the visual technique was easy to understand once the initial layout of the 'Signature' – a grid of coloured values with 7 rows for each day of the week and 96 columns, representing every 15 minutes in the day from 00:00 to 23:45 – had been clarified.

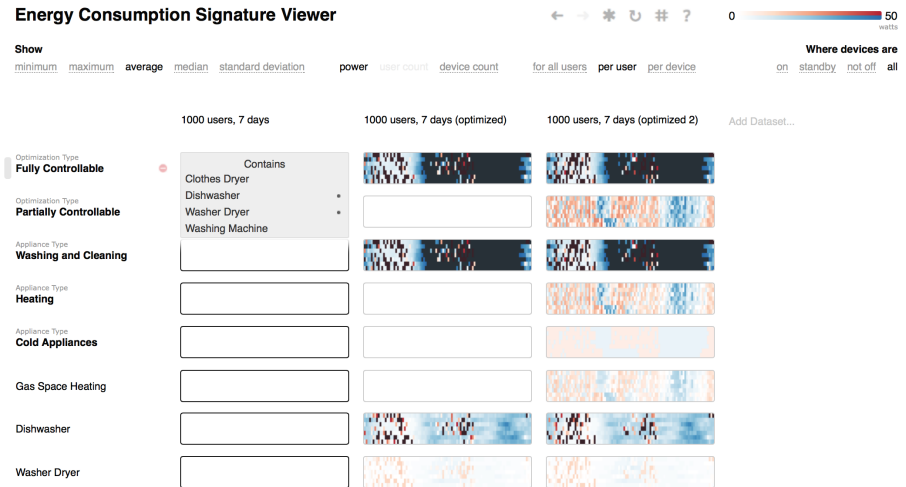
CHAPTER 3. CREATIVE ENERGY DOMAIN EXPLORATION



(a) The 9 different types (A-I) of appliance patterns for the sample dataset for Jan, Mar and May, with a best fit legend and showing the add appliance/group option



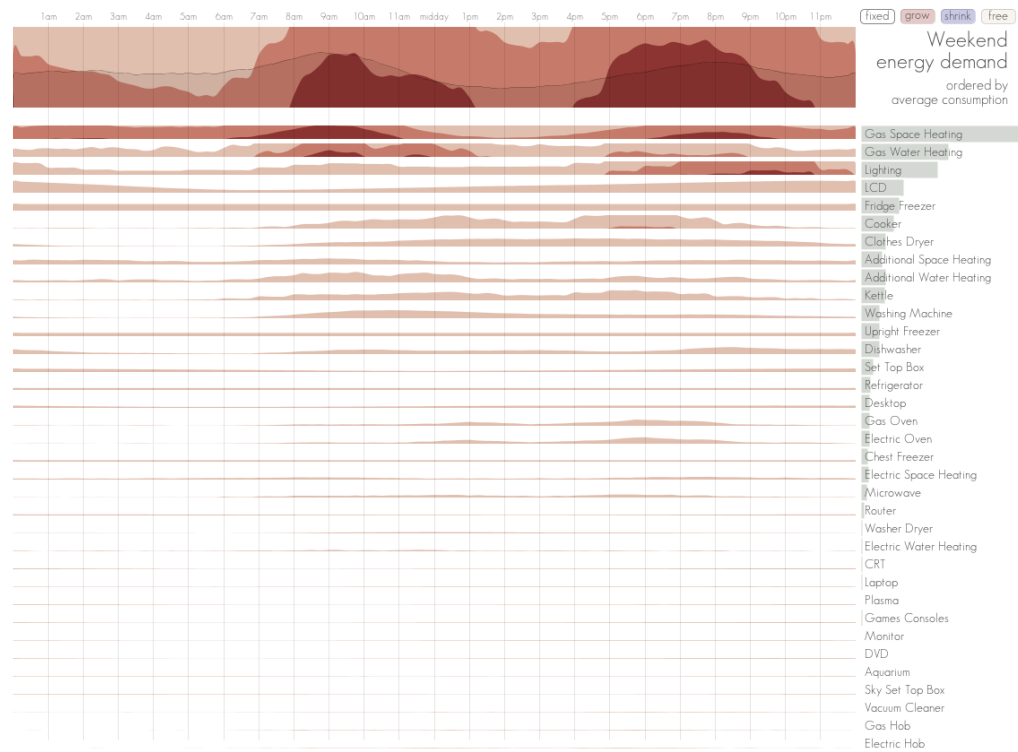
(b) Controllable and non-controllable groups and individual appliances for Jan, Mar and May with 1000 watt legend



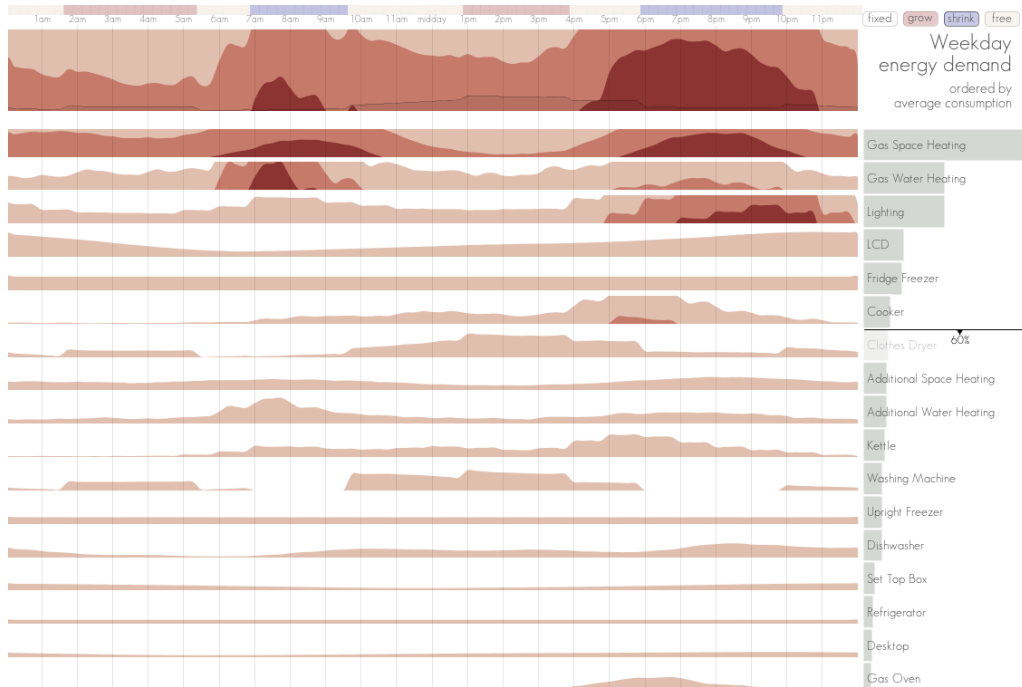
(c) Fully and partially controllable appliances comparing signatures for two types of optimisation algorithms (+/-50 watt legend)

Figure 3.3: Screenshots from Consumption Signatures: Showing average energy consumption as Heatmap ‘Signatures’ representing 7 days (Mon-Sun: vertically) by 15 minute intervals (00:00 to 23:45: horizontally) for each appliance or group of appliances (vertically) by dataset (horizontally)

3.5. EVALUATION: PROTOTYPE DESIGNS, CREATIVITY & PROCESS



(a) Weekday Consumption showing 'gas space heating' as a masked graph on top of the total



(b) Data Sculpting options: Shrink, Grow and Free for shifting Washing Machine and Clothes Drying from Peak to Off-Peak hours

Figure 3.4: Screenshots from Demand Horizons: Showing energy consumption as individual (small) and total (large) horizon graphs for appliances (vertically) by hour of day (horizontally)

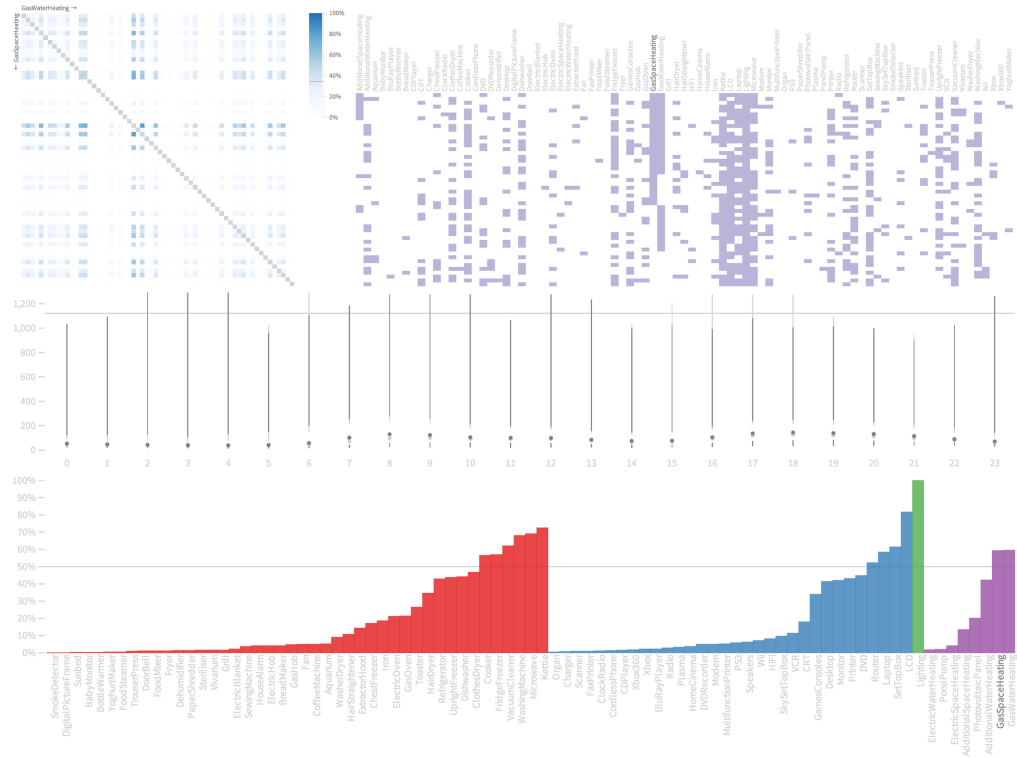


Figure 3.5: Screenshots from Ownership Groups showing total and selected appliance consumption and ownership – box plots show hourly average energy consumption in grey and selected appliance(s) in black, bar chart shows percentage of appliance ownership in sample group currently ordered by type of appliance and ownership and two matrices displaying co-ownership of appliances in different ways

3.5.2.3 Demand Horizons

The horizon charts in this prototype produced many thought provoking ideas once time to understand the technique of horizon graphs had been taken (see Fig 3.4a). The new ‘Data Sculpting’ tool (see Fig. 3.4b and video⁵) produced in the second iteration was seen as very useful and prompted new ideas from both the analysts and the data modellers (see Section 3.5.3). This feature and the creative process leading up to its development is described in detail in the TVCG paper (see Appendix A.1 – Goodwin et al., 2013).

The re-ordering of the horizon graphs for each appliance was also very useful to see how each appliance contributes varying amounts to the total energy use at different times of the day, a fact which was well-known to the participants but they had yet to see visually: *“Very interesting. You just could not get that out of numbers”*. The tool returned a modal score of 2 for the questionnaire responses (see Fig. 3.1) and many uses for the tool were discussed, even a potential customer view as it had such a simple, appealing and engaging design. This demonstrates that a creativity-informed data prototype can open up opportunities that had not been considered in the (creative) requirements workshop.

⁵InfoVis supplementary video available in the digital appendix and online at: <https://vimeo.com/69185134>

3.5.2.4 Ownership Groups

The final prototype was based on the idea of ownership of appliances. The tool scored 3s to 5s in the questionnaire and was not seen as immediately useful by the analysts. While the idea for breaking down the data based on ownership types was seen as very interesting and potentially useful to the market analysis department (rather than specifically the Future Technology Team with whom we were collaborating), the modelled data used for the prototype had little relationship between the appliances and the associated households, so co-ownership was arbitrary and made little sense in terms of expectations when visualised. The project team concluded that perhaps synthesised datasets with credible groupings of appliances should have been used for this prototype.

The design consists of a simple bar chart of appliance by total consumption and box-plots showing average consumption, which are selectable to create ownership groups. Matrix views showing co-ownership were also added in the second iteration, see Fig. 3.5. In terms of design there was also some negative feedback relating to certain aspects being too simplified or abstract and consequently confusing for potential users. The designer had designed to the notion of “*simple*” and “*beautiful*”, using Edward Tufte’s data visualisation design principles (Tufte, 1983), as these had been identified in the Requirements Workshop as being important to the participants. This is a very useful piece of feedback which reminds a visualisation designer that the design itself must match the user’s primary needs and despite requirements, overly simplified graphics may not always be the right choice for the audience or for their data. This also demonstrates the downstream effect outlined in the nested model (Munzner, 2009), that a poor choice made at the abstraction stage effects each of the downstream decisions and will not result in an appropriate solution.

One aspect of this prototype which stood out was simply being able to re-order the bars in a bar chart which represented appliance by proportion of ownership within a given population. After the second iteration it was possible to re-order the bars in five ways, such as by power load or appliance type (see Fig. 3.5), which was new to the participants and revealed some interesting patterns.

3.5.3 Evaluation with the Data Modellers

In addition to the evaluation with the energy analysts, the energy modellers, who had generated the data model used in three of the four prototypes, were also interviewed to informally evaluate the prototypes. The modellers had been engaged throughout the process, although had not been directly involved in any of the workshops or design decisions. As the modellers had in-depth knowledge of the data model, it was seen as appropriate to use their knowledge in the evaluation stage. The data modellers regarded

the designs to also be novel: *“they give me the opportunity to analyse the data in a different way.”* and allowed them to see surprising structure in their data: *“I didn’t expect to see these patterns”* and *“I wouldn’t be able to spot the problem before I saw this graph.”*. Their view on the trial data also changed completely upon seeing the data visualised through the *Smart Home HeatLines* tool: *“before I thought the trial data could not be used due to errors and outliers. The visualization showed me that you can use this data and detect different patterns and user behaviour.”* It was discovered that all four prototypes were very appropriate to the needs of a modeller: *“The way you solve a problem is by doing some visualisation in your mind and these tools help you greatly to facilitate that.”*. Opening up clear opportunities for improving data visualisation within the energy data modelling domain: *“it has got great potential ... to spot problems, abnormalities, see the patterns, come up with new ideas, new theories, new models.”*

3.5.4 Creative Design Process

The structured group discussion during the User Evaluation Session revealed that the analysts felt that the process was engaging and that they had contributed and learnt from the process: *“It has been an enormous learning curve. A great learning curve.”* The process was deemed to be educational and stimulating helping the analysts to understand the possibilities of data visualisation: *“I realise that actually this has got many potential applications and many many uses”* as well as the value of visual design: *“the data is a crucial thing and the visualisation of that data is almost as important to move [...] from information to insight.”*. The analysts were very pleased with the responses to the feedback: *“you actually listened to our feedback, helped us shape that feedback and then delivered”* and the use of parallel prototyping to bring four prototypes instead of one solution. The four prototype designs were described by one of the analysts as *“creative approaches which show us the density, variability and value of our data”*, while the use of different techniques were seen as *“very different”*, new, novel and valuable: *“you have brought something that we couldn’t have thought of ... and the [Smart Home] project will be better for it.”*.

The use of creative activities which built upon the previous during the Requirements Workshop, not only helped to keep the flow of thinking, but allowed ideas to be reiterated and enabled the project team (designers) to gain a rapid overview of the problem domain. The use of creativity techniques seemed to help to *“push domain experts to discuss problems, not solutions”* (Sedlmair et al., 2012, pp.2436). Perhaps this was partly due to the fact that the datasets and possible visualisation methods were largely unknown to the analysts prior to the workshop. The creative activities certainly engaged the analysts and allowed the designers to understand the potential tasks and

domain problems. The Requirements Workshop, along with talks about the model with the data engineers, formed the domain characterisation stage from the nested model of visualisation design (Munzner, 2009). The process continued with the creation of the concept map, requirement feasibility and the Design Workshop, which mapped the problems to the available data and forms the data abstraction stage (Munzner, 2009). Visualisation designs and interaction were encoded as part of the design workshop and continued during the development stage with analyst feedback. The rapid parallel prototyping was useful for tackling many of the problems at once with multiple solutions, following creative design process principles, as described in Section 2.2.2. The user feedback and evaluation sessions allowed each of the prototypes to be assessed for appropriateness and validated by the users. Potential threats (Munzner, 2009) were also acknowledged, for example the fact that the wrong abstraction choice had been made when designing Ownership Groups. The deployment of the prototypes was not part of the research project goals and these prototypes have yet to be developed further. Testing the designs with users in-situ for solving the particular tasks is therefore not yet possible.

To complete the design process evaluation, the designers reflected on the process in comparison to previous projects and concluded that it was more creative than usual. The process is summarised in Fig. 3.6 with the creativity techniques (highlighted yellow) inserted at the early stage of the process, with the intention of introducing a creative climate throughout. The creative activities were seen to help the designers and domain experts communicate, share experiences, establish trust and work together. Creative thinking was experienced throughout the process, not only from the designers but also from the analysts. Some individual aspects of the process were deemed to be more creative than others and these are highlighted in orange in Fig. 3.6. Some individual aspects, in particular the data sculpting feature for Demand Horizons, which allows the user to manipulate the model and simulate the optimisation process, are discussed in greater detail in the TVCG paper (see Appendix A.1 Goodwin et al., 2013). While the use of reflection is subjective, this research begins to evaluate creativity in the design process, which has been expressed as necessary in the literature (as discussed in Section 2.2.4).

3.5.5 Evaluation Conclusions

In conclusion, each of the prototypes were seen as appropriate and creative to varying degrees, with three of the four particularly useful as the analysts could envisage them being used in the company. The simple tasks of being able to re-order and re-shape the data onscreen and see these transitions occurring were seen as very beneficial in all four prototypes. Some degree of familiarity of graphics (as well as data and patterns) was

preferred and the well-designed use of the layout and consistent use of colour was seen as beneficial, particularly when large amounts of data are being shown on the screen all at one time. This indicates that while there is a place for creativity and novelty, the design should not be too abstract and good design principles should be followed. The use of creativity techniques enabled the users to be open to expressing new and potentially novel ideas, while the user feedback was extremely useful for improving and focusing the prototypes during the second stage of development. The use of creativity techniques, rapid prototyping, user feedback and agile methods were all seen as beneficial to the process and are considered for future work in the research for this thesis.

The fact that the creativity-informed data prototypes opened up opportunities, changed expectations, needs and ambitions are all important results of the research, which demonstrate that there is a real benefit for using creativity techniques within visualisation design processes and in particular for identifying the benefit of visualisation for new data sources. The creative process opened up many additional channels of research (Wood et al., 2014) for continued investigation of smart home energy analysis and visualisation. A general outcome of all the prototypes, identified by both the energy analysts and the data modellers, was that the datasets could be visually explored and analysed with greater ease than before, revealing a clear need for improved visual analytics and visualisation for both energy analysts and modellers for current as well as future analysis needs.

3.6 The Who? and Where? of Energy Consumption

The four prototypes described above demonstrate possibilities to investigate the *what?* and *when?* of energy use, but the *who?* and *where?* (and *why?*) were also seen as equally important. The investigation of the *who?* and *where?* of energy consumption links to consumer profiling. A number of opportunities for improved consumer profiling were identified by the analysts during the Requirements Workshop as well as identified in the concept map (Appendix B.4). Of the four activities, the first was highly lucrative for aspirations and ideas, with at least 10 linking directly to profiling users or grouping consumers by typical traits (see Table 3.4). During the subsequent activities there was a growing interest in the analysts' want and need to "*make predictions*", "*find typical patterns and trends*" and to "*slice and dice*" the datasets by lifestyle, age, housing, location or time.

Segmenting the consumer data was again emphasised during the Visualisation Awareness activity where the giCentre's PlaceSurvey (Slingsby et al., 2014a) tool was shown as part of the demonstration and led to a vibrant discussion. It was seen as a very useful tool for use within the company, as the participants could import their own

3.6. THE WHO? AND WHERE? OF ENERGY CONSUMPTION

customer insight data and start to segment the data to investigate patterns in consumer traits and behaviour such as by gender, time, lifestyle, own home and energy type. Key questions such as “*how variable is the data?*” became important. In terms of geography, ideas such as “*compare city A to city B*” were seen as very useful, as was an idea of producing hotspots of geographical areas to show *where* the energy is mostly being used by households or potentially *where* it is being mostly (micro-)generated. The need to “*myth bust*” and gain “*nuggets of knowledge*” were emphasised. Of the five key themes which came out of the workshop, the notion of understanding more about customers’ habits is a key objective, in which *who* consumers are and *where* they are located play an important role. These two aspects of energy consumption prompt continued research for this thesis.

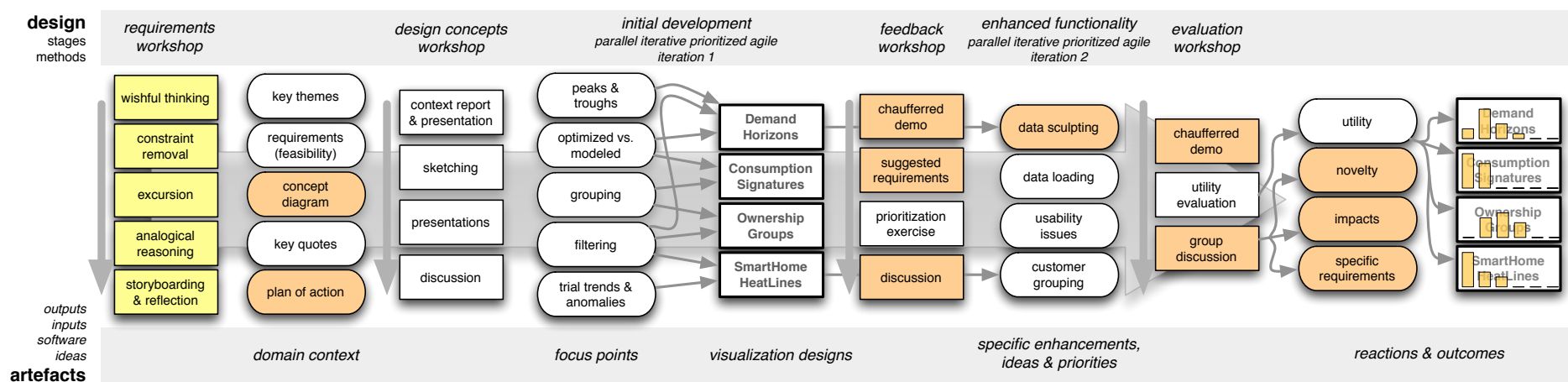


Figure 3.6: The creative design process for the smart home project. Rectangles are techniques (thick edges represent software prototypes). Concepts are round edged. Arrows show direct links between concepts and prototypes. Other links are implicit and less direct. Yellow indicates deliberate creativity mechanisms. Orange highlights processes and concepts in which creativity amongst analysts was strong. Fig. 2 from Goodwin et al.(2013, pp.2619)

I would like to	Aspiration
Know	how lifestyle links to energy use
Know	what does an ‘average’ home do with energy
Know	how electricity and gas use changes with age, lifestyle, life-cycle, group, segment
Know	the who, what, where, when [& why] of energy use
Know	what customer comparisons to use
See	the effect of the technology
See	who is using what and when
See	how comfortable people are
Do	understand how the data relates to the people in the houses
Do	analyse data for insights that are to do with more than the energy used

Table 3.4: Aspirations relating to the segmentation of data by population or lifestyle characteristics

3.7 Chapter Summary

In this chapter the needs of the future and current energy industry are investigated and the smart home project was successful in demonstrating many possible opportunities for the visualisation of smart home data. Three of the four prototypes were deemed appropriate for use and continued development could see these becoming useful tools in the industry. The investigation of the smart home analysis possibilities reveals real need for improved visualisation within the energy sector and a need for grouping and segmenting consumers based on typical traits for better understanding of consumption patterns as well as building customer trust through improved services and knowledge.

In order to begin to tackle the first two of the five key themes identified from the workshop *Analyse the Data* and *Develop Knowledge* (or to tackle the first step of the plan of action: *Discover where energy is going*) four prototype visualisation tools were developed. The four prototypes specifically targeted the *what?* and *when?* of smart home energy consumption, but the *who?* and *where?* were open for continued research due to limitations of the data. With the energy industry facing many changes, there is a clear need for better understanding of energy use and in particular consumer habits and behaviours. As data is only just beginning to become available to industry, any advanced profiling or visualisation needs to be both adaptable and flexible to future needs and open to new and useful datasets. In general, the prototypes produced for the research were seen as very useful for the industry and there is a demand for improved visual analysis for both energy analysts and energy modellers (an unexpected user for the prototypes) who both found useful and unexpected visual insights. The creative process also taught us plenty about inclusive design, its benefits and consequences and the positive role that creativity techniques can play.

4

Exploring Energy-Based Geodemographics

In this chapter, profiling energy consumption based on who people are and where they live is investigated through the context of energy-based geodemographics, as discussed in Section 2.3.2. Section 4.1 describes exploratory data analysis using novel visualisation designs produced using a hierarchical framework (Slingsby et al., 2009, 2010a,b, 2011) for investigating the patterns between energy consumption and geodemographics. The research reveals evidence of differences in consumption both between and within geodemographic groups, indicating that general profiles do not segregate the population for energy profiles and new clusters are needed. This research was presented at the international InfoVis 2012 conference, the national GISRUK 2012 conference and a specialised PhD Symposium on Household Energy Consumption (see Appendices A.3, A.4 and A.6).

The geodemographic generation process is explored through the use of two tools: `gd` package for R and GeodemCreator in Section 4.2.1. The simplified four stage process, as described in Section 2.3.4 (and shown in Fig. 2.3) is used to investigate current visualisation examples as well as outline the proposal for a visual and interactive approach to aid the generation process. Design ideas for the four stages are described in Section 4.2 and were presented at the annual meeting for the North American Cartographic Information Society (NACIS) 2013 (see Appendix A.5). The variable selection stage (Stage 2) is shown to require particular attention for visualisation research and forms the focus for much of the research in the remainder of the thesis. The investigation continues with an exploration

of possible candidate data variables and sources (linking to RQ1) in Section 4.3 for the generation of an energy-based geodemographic classification.

4.1 Exploratory Visual Analysis

This section describes an exploratory visual analysis process which was used to begin to investigate the *who?* and *where?* of energy consumption. Energy consumption data is investigated in combination with two geodemographic classifications. This exploratory analysis continues research from Druckman and Jackson (2008) (described in Section 2.3.2) using a hierarchical visualisation method to investigate high or low levels of energy consumption in connection with geodemographic groups, with the representation of the geographic variation in these relationships. The analysis reveals that some groups have far higher consumption than others, but variation occurs geographically. The investigation indicates that more in-depth research is needed into which variables correlate with energy use and which would produce useful energy user profiles, linking to RQ1.

4.1.1 Data and Limitations

Three datasets are used in this exploratory analysis:

1. Sub-national annual average electricity consumption for 2008 based on ordinary electricity meters, available at LSOA (2001) level and updated annually by the DECC (2008). Gas is also available but as some geographical areas do not have gas pipelines this exploratory analysis concentrates on electricity, as this is nationwide;
2. Experian's MOSAIC Public Sector Classification for 2010, available at LSOA (2001) level for academic research¹, with 15 groups (shown in Fig. 4.1) and 64 Type;
3. Output Area Classification 2001, openly available at OA (2001) level from the ONS (2005) with 7 Super Groups, 21 Groups and 52 Sub Groups (illustrated in Fig. ??).

At the time of this research (2012) the 2011 Census variables and subsequent OAC 2011 had not yet been released and MOSAIC 2010 was seen as the most relevant available data to the current population, as demographics change fundamentally over a ten year period (Gale and Longley, 2013). The variables and methodology for MOSAIC are, however, unpublished and therefore in-depth investigation of variables was not possible. OAC 2001 was investigated together with MOSAIC, to allow for the open variables and the methodology to be investigated if needed. As the DECC and MOSAIC data are both available at LSOA (2001) level, the OAC data was also aggregated to LSOA level by taking the dominant Super Group and Group for each LSOA.

¹Available from www.mimas.ac.uk

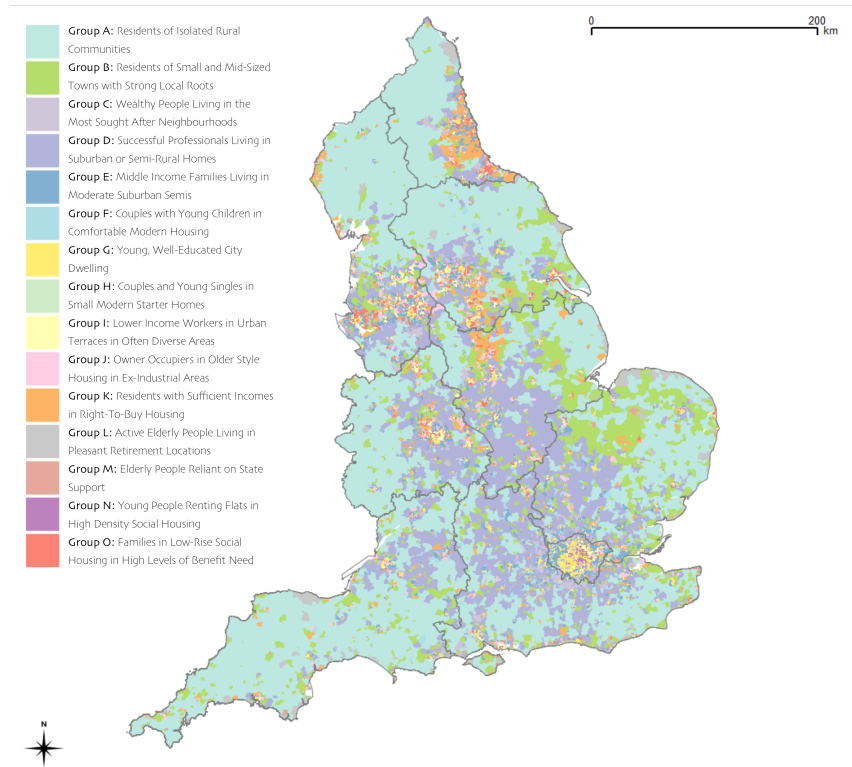


Figure 4.1: The 15 Groups of Experian's MOSAIC Public Sector Classification 2010 showing geographical locations and group descriptions. Contains National Statistics data ©Crown copyright and database right 2012, Ordnance Survey data ©Crown copyright and database right 2012. MOSAIC available via academic license from www.mimas.ac.uk, 2012

The DECC data is classed as experimental as the methodology is still being developed and each year improvements are made (DECC, 2008, 2013a). A number of meters and consumption are ‘unallocated’ to a geographical region, with some regions being allocated too few and others with no data at all. The dataset is therefore not 100% complete to the small-area level; however, it is the only openly available national dataset of energy consumption at this level of detail. The use of areal units of geography is always slightly uncertain as data is aggregated to a small area and social diversity is lost in the aggregation process. The statistical boundaries of OAs and LSOAs used in this research were created with statistical homogeneity in mind and therefore the risk of using areal data at this level is reduced, when comparing this to using areas not designed for statistical homogeneity, such as Wards or Postal units (as discussed in Section 2.3.3).

4.1.2 Visual Technique

Hierarchy plays a key role in this comparison process as there is a need to identify energy consumption compared to the geodemographic groups or types as well as investigate whether geographical location is of relevance to the consumption. As OA and LSOA are statistical boundaries, their names are often not common knowledge and therefore referencing more well known geographical units, such as local authority, town or region

allow for easier comprehension. Traditional choropleth maps are useful for presenting a variable and the geographical regions which they are representing; however, visualising more than two hierarchical levels at one time is difficult, unless multiple maps are shown in juxtaposition. Choropleth maps are also known to mask information about densely populated areas (as discussed in Section 2.4.3). In light of the results of the smart home analysis (Chapter 3) and the need to represent populated as well as rural areas, other creative, novel and alternative visuals, which allow as much information to be reported at one time, were investigated. Due to the need to represent hierarchy, space filling rectangular cartograms (spatial treemaps) were chosen. These examples were produced using HiDE software (www.giCentre.org/hide) and the hierarchical method established within the giCentre (Slingsby et al., 2009). The visuals created appealed to academics in the energy, geography and visualisation domains at different academic conferences.

4.1.3 Visual Analysis

Fig. 4.2 shows three hierarchical representations of household electricity consumption for England where consumption is coloured light to dark. In the first illustration Government Regions of England are represented spatially and contain smaller rectangles for Local Authority regions, also spatially ordered. The size of the rectangles relates to the number of electricity meters in the area (a proxy for number of households as most households have only one meter) and this is consistent through all the visuals. The first image in Fig. 4.2 (Geography), represented only by geography, does not show much variation, except that South, particularly the South East (labelled as SE), has more households and in general more areas with higher consumption than the North. London (labelled as Lon) shows higher consumption in the outer boundary rather than centrally, despite household/meter numbers being relatively similar. This is opposite to the pattern of the East and South East, which show far higher levels in the central areas than the outer areas, particularly the areas which map to the coast in each case. A drawback of the rectangular cartogram representation is that it uses the centroid of the region to map the rectangle to the layout and sometimes this is surprising; for example the location of London is the bottom right corner of the layout as the centre of London is further to the east than that of the actual region of ‘South East’. A useful feature of the HiDE software, over the static visuals shown here, is that it allows interactive switching between actual and rectangular layouts to enable the user to fully understand the representation.

The second representation in Fig. 4.2 (Geodemographic Group) shows a hierarchical layout of 15 MOSAIC Groups containing 64 MOSAIC Types. These are ordered alphabetically as there is no spatial relevance. Firstly, it shows that the MOSAIC Groups are different with Group E, B G, K, I and D containing more households than

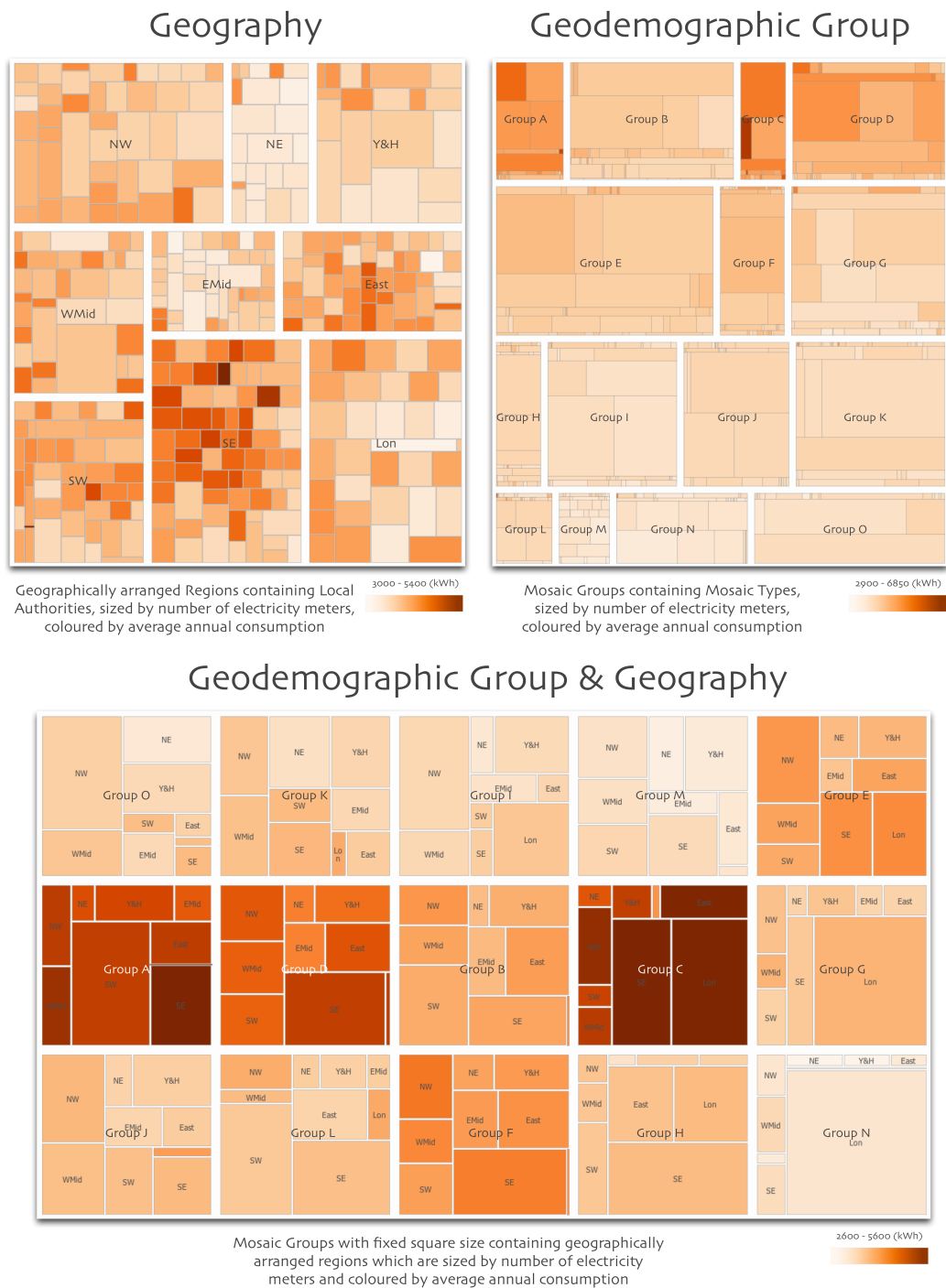


Figure 4.2: Three hierarchical representations of electricity consumption (colour): a. (top left) Government Regions in England containing local Authority areas – ordered spatially, b. (top right) 15 MOSAIC Groups containing 64 MOSAIC Types – ordered alphabetically, and c. (bottom) 15 MOSAIC Groups containing English Government Regions – ordered spatially



Figure 4.3: Hierarchical representation of electricity consumption (colour) by the 7 OAC 2001 Super Groups, containing all Local Authority regions, sized by number of electricity meters and ordered spatially

the others. The top row contains most of the darker rectangles representing the higher consumers, although the colour of Group B remains similar to the Groups in the other rows. Differences within the MOSAIC Groups show variations by MOSAIC Types. In general, Groups A, C and D are shown to be high consumers, with some Types more than others. When the geodemographic names and profiles are noted (see Fig. 4.1), this reveals a correlation of higher electricity use with higher disposable income households (Group C: ‘Wealthy People’) or rural/suburban locations (Group A: ‘Rural Communities’ and Group D: ‘Suburban’). This finding is in accordance with the literature (as discussed in Section 2.1).

The final illustration in Fig. 4.2 (Geodemographic Group and Geography) shows 15 MOSAIC Groups, of equal size for ease of comparison, containing English Government Regions, ordered spatially and again size relates to the number of electricity meters. Here, although clearer than in the other representations, groups A, C and D are darker and a geographical variation within them is evident. The spatial variation of consumption within these groups is of particular interest, e.g. within Group A there is very high electricity use in London, yet moderate use in North East (NE) and East Midlands (EMid). This highlights that energy use varies within the Groups by Type, but also geographically.

The investigation of the OAC 2001 Super Groups and electricity use (see Fig. 4.3) also shows a visual pattern, although not as well defined in this case, as there are only 7 Super Groups rather than the finer segregation of 15 Groups in Mosaic. The Super Groups of ‘Countryside’ and ‘Urban Fringe’ show higher consumption, linked to the ‘rural/urban’ connection, again with very clear geographical variation within the groups.

4.1.4 Summary of Exploratory Visual Analysis

The exploratory visual analysis discussed in this section confirms findings in the literature that certain traits of population are shown to use more electricity than others, for example populations with higher income. Electricity consumption is also shown to vary with geographical location, certainly at the limited scales shown here, with a geographical variation evident within geodemographic groups. The results indicate that the use of such general geodemographic products are not segregating enough for energy use, as the groups or types (of MOSAIC or OAC) are not directly linked to energy consumption. While some similarities can be seen, the geographical clustering of energy consumption data together with relevant variables of household and population characteristics to create an energy-based geodemographic classification could improve consumer profiling and the interpretation of both demographic and geographic variations in household energy usage.

4.2 Exploring a Visual Geodemographic Process

This section uses a collection of the available data variables (1 and 3 in Section 4.1.1) to investigate the available tools for generating geodemographics and the opportunity for a visual and interactive approach to aid the generation process. Possible design ideas for each of the four stages identified in Section 2.3.4 and shown in Fig. 2.3 are investigated. The design ideas begin with an open activity to identify possible requirements and ideas are sketched for possible designs. In particular Stage 2, the variable selection stage, is shown to be of particular need of visual representation due to many complexities and sensitivities associated with the variable decisions and the lack of visual representation of these in the current tools.

4.2.1 Investigation of the Available Tools

In order to investigate the geodemographic classification process, firstly the open source software environment R (R Core Team, 2014)² was used, together with the *gd* (geodemographic) package by Singleton (2012). For this investigation the variables in Section 4.1 were used: the 41 OAC 2001 variables, together with the 2008 electricity and gas consumption variables from DECC.

The *gd* package aids the creator in dealing with some of the complex processes of geodemographic data processing and generation of clusters. The raw data is linked with a transformation specification file which is used to transform each variable. This includes whether the column is a numerator, denominator or a non-count variable, whether index scores or percentages are required for the count data and what type of algorithm is required

²<http://www.r-project.org>

4.2. EXPLORING A VISUAL GEODEMOGRAPHIC PROCESS

to normalise or transform the data. Using `gd` two options are available: Log or Box Cox. Reflecting the literature of previous geodemographics examples (see Section 2.3.3.4 for the discussion on k-means and alternative clustering methods), the package uses the k-means clustering algorithm for the clustering process. The package provides static visuals to compare the correlated variables (as shown in Fig. 2.5) and further graphics allow the creator to investigate the sum of the square means for the k clusters and decide on the number of k to choose.

GeodemCreator (Adnan, 2011) was also investigated. Although not openly available yet, the tool was created for PhD research in 2011 and was obtained from the author for investigation. This tool allows users to create their own bespoke, domain-specific or local classifications in real time. The tool was developed using Java and R and therefore can run on any operating system. It has two modes; basic and advanced. Basic mode is for inexperienced users *“who want the software to create a classification for them without the need to specify variables, their weightings, the number of geodemographic classes (groups), or the choice of clustering algorithm”* (Adnan, 2011, pp.209). Like the `gd` package, the tool uses k-means for the clustering algorithm. This is due to speed of calculation as the tool is designed to build bespoke classifications in real-time (Adnan, 2011). The basic mode allows users to build local or domain-specific classifications based on the 41 OAC 2001 variables. Users can define the local geographical region for the classification and/or the five domains of OAC (or a general classification based on all five), but many decisions including the number of k clusters are pre-determined; k is based on the within-sum-of-squares values for the default datasets.

The advanced mode allows users to have greater control; Users can specify individual variables, upload and use their own variables, choose the number of clusters and define individual variable weightings. This, like the `gd` package, is expected to be used by expert users who understand the procedures and methods of creating a geodemographic classification. In terms of correlation, the software provides the ability to test the correlation of the uploaded variables and a warning is given to the user for highly correlated variables. Consistency of all input variables across the same geographical extent is also highlighted by the software and a warning message is given if the uploaded variables do not match the pre-loaded OAC 2001 variables. The user can choose how many k to use based on the within-sum-of-squares plot provided. In terms of output, GeodemCreator produces a CSV file with a cluster number assigned to each OA. Radial diagrams are also produced for visual analysis and interpretation of the cluster profiles. Adnan (2011) explains the use of the tool and these diagrams in a number of case studies based on different geographical areas and subsets of data. This tool, when openly available, could greatly benefit those who want to profile customers or local populations.

Although the two tools aim to aid expert users in creating bespoke geodemographic classifications many open questions remain, such as: *‘how many clusters to use?’*, *‘what is the effect of removing a variable?’*, *‘how does the scale of the data affect the results?’*, *‘does transformation improve the results?’* and if so *‘should all variables be transformed in the same way?’*. The decisions made during the process are subjective to the creator as well as limited by the constraints of the tools. As expressed in the review of the literature (Section 2.3.3), finding a stable, reliable and valid clustering result takes much time and to ensure reproducibility of the process a lot of documentation about these decisions is needed. The fact that the process is so time consuming and intensive hinders the ability for novice or middle-ground users (those without the expertise in the processes) to create domain-specific classifications. It is argued therefore, that a more visual and interactive approach to generating geodemographics would enhance the process and could aid the understanding and uncertainty of the complex statistical methods used, such as transformations and clustering.

4.2.2 Requirements for Visual Representations

To investigate this problem in greater depth, one of the methods for investigating new ideas and stimulating creative thinking from the smart home requirements workshop was utilised. The method used for the wishful thinking activity in Section 3.2, using the questions adapted particularly for data visualisation; *‘What would you like to know?’*, *‘What would you like to be able to do?’* and *‘What would you like to see?’*, was used by the author in order to build requirements for a potential visualisation tool to aid the generation of a geodemographic classification. This technique was used as it had been shown to be particularly beneficial in the requirements workshop with the analysts in Section 3.2.

As there are only few experts in geodemographics, acquiring users for a creative workshop to build a list of requirements or aspirations was not possible for the research. Rather than relying on non-expert users, the requirements generated here were based on a thorough literature review (Section 2.3), the testing of the current tools (gd and GeodemCreator – see Section 4.2.1) as well as informal consultation with Chris Gale (who was creating OAC 2011 at the time) (Gale, 2014a). Having investigated the literature (Section 2.3) and experimented with the process (Section 4.2.1), the wishful thinking activity was carried out by the author in order to think creatively about *‘the visual and analytical possibilities for the process of generating a domain-specific geodemographic classification’*.

The activity resulted in many individual statements, which were then grouped into topics. Topics included statements relating to: clustering methods or clusters, correlation,

distribution (both spatial and statistical), data scale, statistics, transformations, visual layout, methods of interaction and use of colour. The wall of statements are all shown in Appendix B.7. Many of these can be expressed as user stories (US) for the role of an energy analyst creating a geodemographic classification, i.e. US#1 can be expressed as: *“As an analyst, creating a energy-based geodemographic classification, I would like to see the distribution of the variable, so that I know if it is heavily skewed”*. These user stories are shown in Table 4.1. They are referred to by number when investigating visual ideas in the following section and referenced when building and evaluating the framework instantiation in Chapters 6, 7 and 8.

4.2.3 Visualisation Examples and Ideas

The knowledge of the process, testing of other tools, the user stories and statements discussed in the previous sections and the review of the literature (Section 2.3) were all used to help investigate design ideas for each of the four stages of the process (as described in Section 2.3.4). Each stage has already been illustrated to varying degrees in visualisation in prior research (as discussed in Section 2.3.5). Singleton (2007) investigates the anomalies when combining the geographical regions, which could be classed as a useful visual for Stage 1. The gd package (used in Section 4.2.1), assumes that Stage 1 and 2 are completed but assists with the running of clusters with visual graphics (although static) for Stage 3 and Stage 4. Stage 4 is visualised statically in the GeodemCreator (Adnan, 2011) where radial diagrams are used to distinguish classes. OAC 2001 and 2011 also use radial diagrams and other static visuals for validating and understanding the results of Stage 4 (Vickers et al., 2005; Gale, 2014b).

In terms of interactive visualisation there are a number of examples for classification and clustering as discussed in Section 2.3.5 and 2.4. The verification process of Stage 4 is investigated by Slingsby et al. (2010b), the clustering process of Stage 3 has many examples (e.g. Cao et al., 2011; Choo et al., 2010) and the variable selection process of Stage 2 is investigated in detail by Seo et al. (2002; 2005), although this is not specifically for geodemographics as geography is not included.

It is evident that developing an application to fully assist the generation of a geodemographic classification would extend current work in domain-specific geodemographics (Singleton, 2007) and real-time geodemographics (Adnan, 2011), with the addition of visual aid and visual comparison, which draws upon research in the visualisation community on visualising clustering (Cao et al., 2011; Choo et al., 2010), geodemographics (Slingsby et al., 2010b, 2011) and multi-variate comparison (Seo and Shneiderman, 2002, 2005; Turkay, 2013). Initial ideas and designs for a four stage visual, iterative approach to generate geodemographic classifications is shown in Fig. 4.4 and

US#	I would like to...	Wish (do/know/see)	Reason (so that I can ...)
1	See	the distribution of the variable	know if it is heavily skewed
2	See	as many variables as possible	get an overview of them
3	Know	which variables are geographically 'interesting'	create better clusters
4	See	variables by domain	quickly identify them
5	See	structure in the multivariate data	gain insights and knowledge
6	See	pairwise correlation	see heavily correlated variables
7	See	highly correlated variables	decide whether to include them
8	Do	remove variables from view	concentrate on specific ones
9	Do	highlight important variables	quickly identify them later
10	Do	reorder variables systematically	see which are similar
11	See	variables duplicating the same information	decide which to remove
12	Do	update variables when new data is available	keep the classification up-to-date
13	Do	add the smart home variables	know how useful this is in the future
14	See	how local statistics differ from regional and global	see geographical patterns within the variables
15	See	the local statistics mapped	see geographical variation within the variables
16	See	switch between geographical and statistical view	understand the geographical variation
17	Know	when scale effects correlation	know which variables are sensitive to scale
18	See	the results at different scales	know which scale is appropriate for my variables
19	Know	the variable denominators	know how the data was constructed
20	See	the original data	see how it is transformed
21	Do	apply different transformations	see the effect
22	See	both transformed and 'normal' data	see the difference
23	See	the proportion and significance of outliers	know whether to exclude them
24	Do	apply threshold values to variables	remove outliers
25	Know	which variables to look at first	know where to start the investigation
26	Know	the original variable values	understand the effect of standardisation
27	See	<i>the changes (e.g transform) occurring on the screen</i>	<i>know what the method is doing</i>
28	Do	filter by geography/attribute/time	concentrate on one area/domain/period
29	Do	aggregate by geography/attribute/time	see aggregated patterns
30	See	comparison across multiple scales	see the effect of scale
31	Do	select variables interactively	compare them
32	Do	group/combine variables	see if this makes them less correlating
33	Do	split up combined variables	see if this reduces correlation
34	Know	how robust a variable is	trust it/use it
35	Know	how the energy variables effect the clusters	know which variables to include in the clustering
36	Do	adjust weightings	see the effect of the variables on the clustering
37	Do	remove variables from the analysis	improve the clustering results
38	See	the within-sum-of-squares	compare cluster results
39	See	how the variables effect the clusters	know whether to include them or not
40	Do	run clustering on different scales	see how aggregation effects results
41	Do	run the clustering on the fly	see the results in real time
42	Do	test different clustering algorithms	know which one to use for my variables
43	Know	the cluster uncertainty	know how certain each cluster is
44	See	the clusters mapped	see how they vary geographically
45	Know	how many areas are in each cluster	know that they are evenly distributed
46	Know	how uncertain the cluster results are	know how certain the area is in that cluster
47	Do	merge clusters	improve the classification
48	Know	<i>how the visuals related to the data</i>	<i>better understand the representation</i>
49	See	<i>colour used consistently</i>	<i>interpret the visuals easily</i>
50	See	<i>the data represented in different views</i>	<i>see different types of patterns</i>

Table 4.1: User stories for visual and analytical possibilities for the process of generating geodemographics created from the know/do/see activity. 1-34 in bold are relevant to variable selection

described in the following section. These ideas were presented at NACIS in 2013 with further sketches available in Appendix A.5.

4.2.3.1 Four Stage Design Sketch Explanation

In the design sketch shown in Fig. 4.4, Stage 1 represents visual options for combining and merging datasets. Initially the extent of the data sources are compared and reflected in the map. The Census data is available for the whole of the UK, yet the DECC consumption variables are available for England and Wales and the ‘fuel poverty’ variable only for England. The scale of resolution of data can also be visualised; showing the differences between the geographies of LSOA 2001 and LSOA 2011. Stage 2 shows variable comparison for the selection process with options to keep, drop or merge shown next to the correlation matrix. At this stage the pairwise correlation, variable distribution and geographical variation is important. Therefore, correlation is shown through scatterplots and correlation matrices, with histograms for distribution and miniature maps to represent geographical variation – clicking on the items could enlarge them or show another (spatial/statistical) view. In this sketch, Stage 3 shows a choice of three different clustering algorithms, with options for the number of clusters to create, showing the within-sum-of-squares as well as a scatterplot representing the clusters using colour. Variable weightings are also adjustable on the right. Stage 4 shows cluster results for the total (England) and a subset of the full geographical extent (London) with uncertainty of clusters highlighted in the parallel plot and the clusters ready for naming. At this stage, tests can be run on variable sensitivity and reproducibility. The user can then continue to refine the classification or finalise them and apply the results.

Stage 2 of Fig. 4.4 and an enlarged version in Fig. 4.5 show initial sketches for a possible interactive view for Stage 2, where there is an emphasis not just on comparison of correlation between variables (like is currently the norm in variable comparison for geodemographics see Fig. 2.5), but to allow the comparison of geographic variation of both correlation and variable distribution, in order to visualise how each variable varies geographically. It is evident from the geodemographic and visualisation research that this stage has a lack of research specifically for geodemographic variable visualisation, in particular the representation and investigation of scale and geography. Reflecting on the requirement statements/user stories, there are many which relate to the variable selection stage (see user stories in #1-34). In particular with reference to the geographical distribution of the variable, data scale, uncertainty and the transformation process. While investigating this process, it became evident that the uncertainties and sensitivities related to varying data scale and geography in multivariate comparison form an interesting and complex parameter space. There is limited research investigating the

visual comparison of multivariate data in combination with geography and scale and therefore a research gap to fill. While it is not possible to create the whole visual process it is evident that Stage 2 is in need of continued research. It was at this point in the research, that RQ2 was adapted from *“how can visualisation aid the process of generating a geodemographic classification”*, to *“how can visualisation aid the variable selection process”* and the additional two research questions (RQ3 and RQ4, see Section 1.2) were added to the research goals of the thesis. These three questions, along with RQ1, are investigated in the research which forms the remainder of the thesis. The sketches for Stage 2 described in this section form the start of the ideas for the prototype built in Chapter 6.

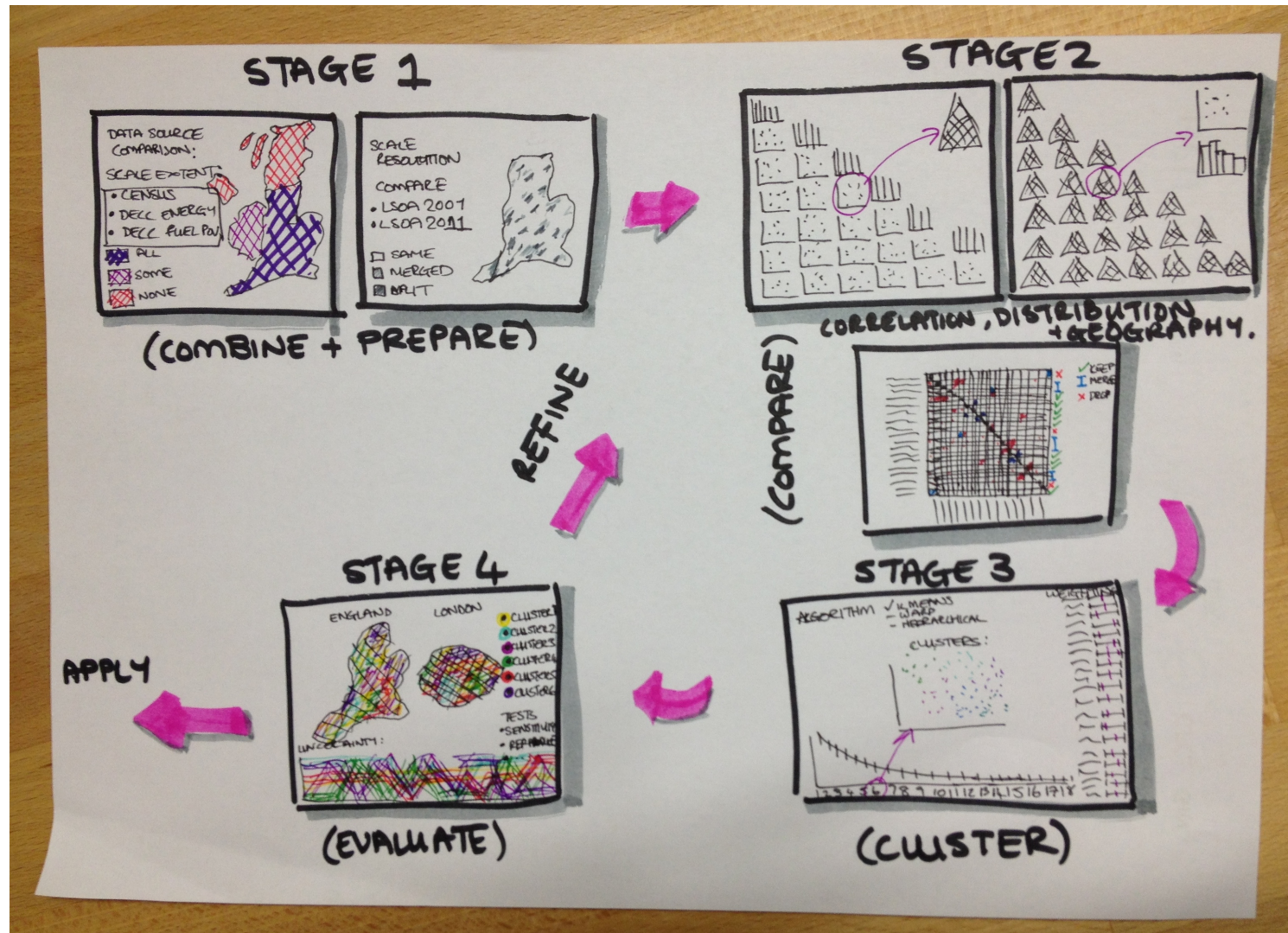


Figure 4.4: Design sketch for a four step application for generating geodemographics with the aid of visualisation at each stage

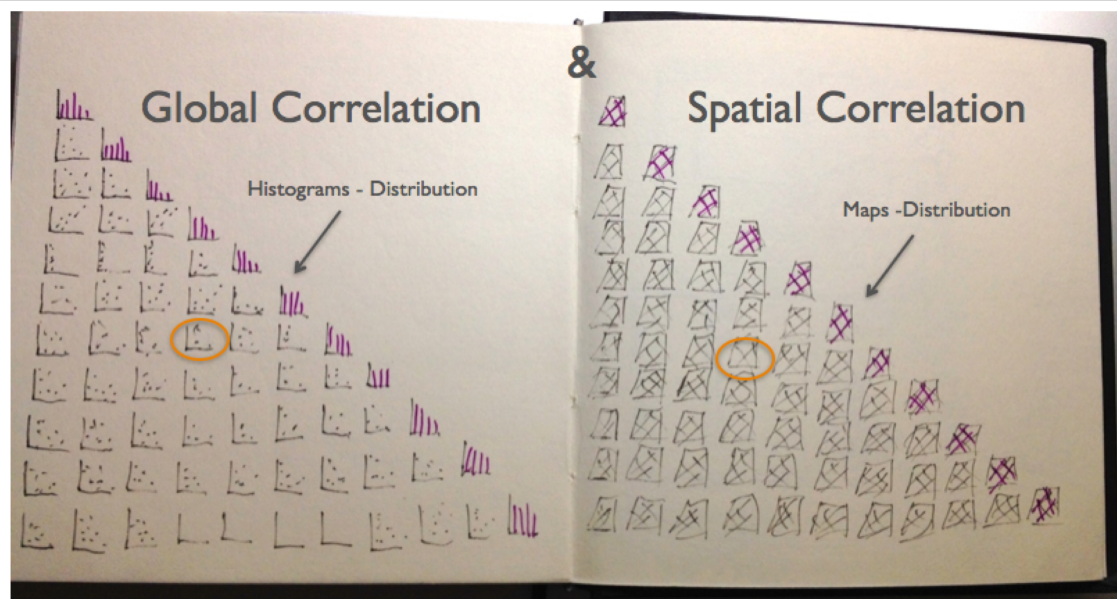


Figure 4.5: Initial Design Sketch for Stage 2: Variable Selection

4.2.4 Summary of a Visual Geodemographic Process

In this section the process of generating a geodemographic classification is discussed, investigated using the simplified four stage process described in Section 2.3.4 and illustrated in Fig. 2.3. The current tools are investigated using demonstrative data and a visual and interactive tool is proposed in order to improve the understanding of the complex, intensive process. There are many options for the visual representation of each of the four stages. Some stages have a significant amount of previous research and there is an abundance of visualisation possibilities. Stage 2, the variable selection stage is seen as most important for further investigation, as geography is a fundamental aspect of the process but is currently only minimally investigated. Continued investigation of the visualisation of multivariate comparison across scale and geography is shown to be important to the process and to academic research.

4.3 Candidate Energy Variable Options (for the UK)

Although data has become more available in the open domain in recent years, there are still limitations on data availability and accessibility. This section identifies possible candidate variable sources for an energy-based geodemographic classification and describes possible future additions based on technological movements in the industry.

4.3.1 Electricity and Gas Consumption

Two variables of particular interest to household energy consumption analysis and classification are electricity and gas consumption, as these two fuel sources are particularly common for domestic energy usage in the UK. DECC collect and publish

4.3. CANDIDATE ENERGY VARIABLE OPTIONS (FOR THE UK)

electricity and gas estimates for England and Wales at LSOA level annually (DECC, 2013a). The dataset contains three electricity kilowatt hour (kWh) calculations relating to ordinary meters, economy 7 meters³ and the total of both types and a fourth variable reports the total kWh of gas. The individual counts of all three types of meter are also included for each area. This dataset is ideal for classification, as it is at a small neighbourhood level (LSOA) for the whole of England and Wales and therefore useful for detailed national profiles. The annual release of the data ensures the data is up-to-date and is also useful for data comparison with other variables. The 2011 release of the data, collected between the end of 2010 and early 2012, make it highly compatible to use with Census 2011 demographic variables, which were surveyed nationwide on the 27th March 2011 (point-in-time collection). While both are annual averages, the individual variables have slightly different collection dates; the gas consumption data spans an annual period (corrected for weather) from October 2010 to October 2011 (collected by xoserve and independent gas transporters), while the electricity consumption data the annual period from January 2011 to January 2012 (collected by data aggregators on behalf of electricity suppliers) (DECC, 2013a). The electricity data from this source for 2008 was described and used in the exploratory analysis in Section 4.1.

A drawback of the 2011 variable is that it was originally released (at the time of this research) at LSOA 2001 boundaries, rather than the LSOA 2011 boundaries like the 2011 Census data. This makes spatial compatibility difficult. Pre-processing was required in the research discussed in Section 5.4, in order to make the variables comparable⁴. The dataset is also classed as experimental as the methodology is still being improved. Some data is ‘unallocated’ to any geographical area (as mentioned in Section 4.1) allocated to multiple LSOAs or has only been accurately allocated to the MSA⁵ level. The unallocated data can either be removed from the classification or split/merged to relevant LSOAs where multiple LSOAs are identified.

While the DECC data is nationwide and at neighbourhood level, the variable is an annual average and therefore temporal fluctuations in electricity and gas use can not be investigated. More detailed data in terms of time and resolution is usually available to analysts within the energy industry, such as quarterly meter recordings or potentially real-time usage data from smart meters (as explained in Section 2.1.2 and Section 3.3); however, customer privacy then becomes a concern, as is the lack of 100% national coverage. The

³Economy 7 is the name of a differential tariff provided by UK electricity suppliers that uses base load generation to provide cheap off-peak electricity during the night. These households require a special meter which provides two different readings. More details: <http://bit.ly/1F94VkO>

⁴Since this research, DECC has re-released the 2011 data using the 2011 LSOA boundaries to make them comparable with the 2012 release of the same dataset and easier to combine with 2011 Census variables (DECC, 2013a).

⁵The 3rd Tier Census Geography with regions of between 2-6,000 households, built up of multiple LSOAs: <http://bit.ly/16XucgG>

DECC data will be used for this analysis as it is openly available, released at a geographical level comparable (although some pre-processing is needed) with the Census data and is available nationwide for England and Wales. When exploring the geodemographic profiling process, its flexibility will need to be considered for allowing future datasets to be added, which may relate to time and appliances, as well as consumption and location.

4.3.2 Other Fuel Consumption and Central Heating Types

Other sources of energy such as solid fuel, oil and bioenergy also contribute to UK domestic energy consumption, particularly for space heating (DECC, 2013d). Limited availability of data for these fuels, particularly at the small-area geographies required for clustering at the neighbourhood level, makes using these variables in the classification difficult; however, gas is the most dominant heating fuel source within UK households with 85% of households in 2011, compared to 8% electricity, 4% oil, 1% solid fuel and 2% other (DECC, 2013d). As gas provision is geographically related to households connected to the gas pipeline network it is important to indicate which areas have high numbers of households using other types of fuel source for their heating as the majority of household energy consumption accounts for space heating. For these non-gas households either the electricity consumption will be much higher than usual (heated by electricity) or the DECC variables will not include the heating consumption due to the use of other fuel sources. To take into account the type of fuel used for space heating the Census 2011 QS415WE data table⁶ can be used. This includes the number of households with no central heating, gas central heating, electric (including storage heaters) central heating, oil central heating, solid fuel (e.g. wood or coal) central heating, other central heating and two or more types of central heating. Since the research for this thesis began, DECC have published detailed LSOA level data for areas with no gas pipeline (DECC, 2013b), which could be used as an alternative; however, the Census fuel type data is at a finer resolution.

Another fuel source to consider is petroleum use for domestic purposes; however, detailed data is difficult to obtain as use is not billed on the household level, like electricity or gas, as this consumption is used for transport, rather than the running of the house. DECC have published Local Authority level data for such residual fuel domestic consumption (DECC, 2012); however, this does not give detailed neighbourhood comparisons and is therefore difficult to consider for a local level classification. Nevertheless, demographic variables from the Census which related to domestic transport use could be included in the classification such as variables related to the number of cars and using public transport for travelling to work. Such relevant variables are included in both OAC 2001 and 2011 (Vickers et al., 2005; Gale, 2014b).

⁶<http://www.nomisweb.co.uk/census/2011/qs415ew>

4.3.3 Energy Ratings

Other energy related variables to consider are energy ratings. The Standard Assessment Procedure (SAP) rating for evaluating the energy efficiency of homes, which has been used since 1993; the higher the rating, the better the efficiency and lower the energy costs (Palmer and Cooper, 2011). Such data is currently only available for new homes at the national level, although national surveys, e.g. the English Housing Survey⁷, are run yearly to include older buildings and statistics are published at the local authority level. As neither data source is a complete picture, nor at the neighbourhood level, it is not possible to include such statistics in a national classification at present.

4.3.4 Heat Loss and Insulation

Heat loss parameters including insulation in combination with efficiency of boiler systems are two important determinants of household space heating use. Both related to the age of the home or investment made by the home owners, as home owners, local authority or social landlords are more likely to invest in energy-efficiency technologies than private landlords (Palmer and Cooper, 2011); however, such variables at small-area geographies are not currently available for analysis. The loss of heat is also dependent on the temperature and weather across the country. Weather data is available for use for weather station locations and could be interpolated across the regions, e.g. from the UK Met Office⁸. Extreme weather conditions tends to affect consumption of space and water heating; however, as this is mainly gas use (85% of households in 2011 (DECC, 2013a)) and the DECC gas consumption variable has been corrected for weather (DECC, 2013a), it may counter affect this correction. If consumption values were available at finer temporal intervals (e.g. daily values) the inclusion of weather would be useful to identify patterns in use, for example consumption on days which are warmer or colder than average. Temperature also links to solar gain potential (also linked to roof size, type and slope direction), and this would be a useful variable to include to distinguish clusters of potential solar generation possibilities. This data is currently not openly available.

4.3.5 Fuel Poverty

Fuel Poverty is a term used by government to define households where the household income is below the poverty line and their energy costs are higher than is typical for their household type (DECC, 2013c). These households have been of particular concern for a number of years (Boardman, 2004) and including such an indicator in this classification could help to highlight areas of fuel poverty related to other socio-demographic variables.

⁷<https://www.gov.uk/government/statistics/english-housing-survey-2012-energy-efficiency-of-english-housing-report>

⁸<http://www.metoffice.gov.uk/public/weather/climate-historic>

Including this in a classification may allow continued research in the area by helping to target certain areas and populations to improve the energy rating of the building and thus help reduce energy bills. DECC provides a variable for ‘fuel poverty’ available at LSOA level for England (not Wales or the rest of the UK) containing the percentage of households which are currently identified as ‘fuel poor’. The 2011 figures are again classed as experimental and are derived from survey data along with fuel prices and income from 2010 and 2011 (DECC, 2013c). The actual survey variables are not disclosed and therefore it is possible that the variable is strongly correlated with other demographic variables; however, it may help to discriminate energy-based clusters. As the variable is available at the same aggregation level as the DECC variables it is seen as useful for further investigation.

4.3.6 Household, Demographic and Socio-Economic Variables

As reported in Chapter 2, there are many socio-economic variables which show correlations with energy use, in particular: household income, dwelling type and size, property tenure, household composition and rural/urban location. As the generation of a geodemographic classification encourages the use of a large number of variables and variables need to be investigated for skewness, strong correlations and geographical distribution, further investigation is needed to reveal which variables should be clustered together with available energy variables mentioned above.

In Section 4.1 Experian’s MOSAIC was used for the exploratory analysis as well as OAC 2001, as OAC 2011 was not yet available. For candidate variables MOSAIC is of limited use as it does not publish the variables for the classification, although the Census remains one of the main data sources (Experian, 2014). The Census is the main source for small-area neighbourhood statistics in the UK and the variables chosen for the OAC 2001 or OAC 2011 are potentially useful for an energy-based geodemographic as they include many relevant variables such as dividing types of houses into detached, semi-detached and terraced, and the population into age groups. There are five domains covered in both OAC 2001 and 2011: Demographic Structure, Household Composition, Housing Type, Socio-Economic and Employment (Vickers et al., 2005; Gale, 2014b). All of these could potentially influence energy use and therefore may be useful to cluster with energy variables. Most of the original 41 variables for OAC 2001 were retained for OAC 2011. Some variables related to age and employment groups were split, other variables were dropped and additional variables were added to increase the discrimination between clusters.

There are 71 unique variables when combining those chosen for OAC 2001 and OAC 2011 and only one of these is specifically referring to energy use. This variable is ‘% of

households with no central heating’ which was actually used as an indicator of poverty in OAC 2001 and was dropped in the OAC 2011 as values are very low across the country. An additional 72nd variable ‘average rooms per household’ was also considered for OAC 2011 at the time of this research, yet was subsequently dropped for the final classification. For the purpose of the research in this thesis, these 72 census variables are identified as useful to compare to the energy (fuel poverty and consumption) variables from DECC and the central heating type variables from the Census, to indicate whether energy consumption relates to geography as well as demographics. The inclusion of both the OAC 2001 and 2011 variables could also allow for the comparison of the overlapping and duplicating variables. The chosen variable combination may contain additional variables once the initial variable selection begins and these are likely to be very different to either choices for OAC 2001 or 2011, as OAC is a general geodemographic profiler rather than domain-specific.

4.3.7 Candidate Energy Variable Options Summary

There are many potential energy variable sources, which could be considered for energy-based geodemographics, yet many of these are either not openly available to the public (or for academic research) or are not available at detailed resolutions to demonstrate geographical variation. From those available, the consumption variables for electricity and gas from DECC at LSOA level are seen as appropriate for this research, along with the ‘% in fuel poverty’ variable and the four central heating fuel type variables. These variables are combined with demographic variables from OAC 2001 and OAC 2011, totalling 78 (‘no central heating’ is in OAC 2001 and is one of the four central heating fuel types) variables for comparison. These 78 are shown in Table 4.2 with the domain of the variable, the Census table source and the IDs for OAC 2001 and 2011. The new or amended variables used in OAC 2011 offers overlapping and inverse variables to be also investigated. These potential candidate variables are used to describe the complexities of scale and geography in the following chapter and subsequently used in the visual prototype described in Chapter 6.

4.4 Chapter Summary

The exploratory research in Section 4.1 confirms that socio-economic, demographic and locational characteristics correlate with electricity use and that the current geodemographics are not suitable for detailed energy profiling. An exploration of the current tools demonstrates that there are many open questions, particularly related to uncertainty of the process, techniques and algorithms. Given the complexity of the process and the need to be adaptable and flexible as new datasets are becoming available, a visual and interactive tool to help create geodemographics is proposed and

CHAPTER 4. EXPLORING ENERGY-BASED GEODEMOGRAPHICS

Domain Name	Name	OAC2001	OAC2011	Data Type	Denominator
Energy	Central Heating: Electricity			Percentage	Total Households
Energy	Energy: Annual Electricity Consumption			Average	Total Households
Energy	Energy: in Fuel Poverty			Percentage	Total Households
Energy	Central Heating: Gas			Percentage	Total Households
Energy	Energy: Annual Gas Consumption			Average	Total Households
Energy	Central Heating: None	v21		Percentage	Total Households
Energy	Central Heating: Other (Wood, Coal, Oil)			Percentage	Total Households
Demographic Structure	Household: No English Language		v23	Percentage	Total Households
Demographic Structure	Population: Aged 0 - 4	v1	v1	Percentage	Total Population
Demographic Structure	Population: Aged 5 - 14	v2	v2	Percentage	Total Population
Demographic Structure	Population: Aged 25 - 44	v3	v3	Percentage	Total Population
Demographic Structure	Population: Aged 45 - 64	v4	v4	Percentage	Total Population
Demographic Structure	Population: Aged 65+	v5		Percentage	Total Population
Demographic Structure	Population: Aged 65 to 89		v5	Percentage	Total Population
Demographic Structure	Population: Aged 90 and over		v6	Percentage	Total Population
Demographic Structure	Born: in new (post 2004) EU Countries		v21	Percentage	Total Population
Demographic Structure	Born: in old (pre 2004) EU Countries		v22	Percentage	Total Population
Demographic Structure	Born: Outside the UK	v8		Percentage	Total Population
Demographic Structure	Born: United Kingdom and Ireland		v20	Percentage	Total Population
Demographic Structure	Ethnicity: Arab or other ethnic groups		v19	Percentage	Total Population
Demographic Structure	Ethnicity: Asian: Bangladeshi		v16	Percentage	Total Population
Demographic Structure	Ethnicity: Black African, Caribbean or Other	v7	v18	Percentage	Total Population
Demographic Structure	Ethnicity: Asian: Chinese and Other		v17	Percentage	Total Population
Demographic Structure	Ethnicity: Asian: Indian		v14	Percentage	Total Population
Demographic Structure	Ethnicity: Indian, Pakistani or Bangladeshi	v6		Percentage	Total Population
Demographic Structure	Ethnicity: Mixed ethnic group		v13	Percentage	Total Population
Demographic Structure	Ethnicity: Asian: Pakistani		v15	Percentage	Total Population
Demographic Structure	Ethnicity: White		v12	Percentage	Total Population
Demographic Structure	Urban: Population Density	v9	v7	Ratio	
Household Composition	Household: with no children	v14	v25	Percentage	Total Households
Household Composition	Household: with non-dependant children	v15	v24	Percentage	Total Households
Household Composition	House: Communal Establishment		v8	Percentage	Total Households
Household Composition	Household: Full-time student(s)		v26	Percentage	Total Households
Household Composition	Household: Lone Parent	v13		Percentage	Total Households
Household Composition	Status: Married or civil partnership		v10	Percentage	Total Population (16 and over)
Household Composition	Status: Separated/Divorced	v10	v11	Percentage	Total Population (16 and over)
Household Composition	Status: Single		v9	Percentage	Total Population
Household Composition	Household: Single person (not pensioner)	v11		Percentage	Total Households
Household Composition	Household: Single pensioner	v12		Percentage	Total Households
Housing Type	House: Detached	v19	v27	Percentage	Total Households
Housing Type	House: Flat/Apartment	v20	v30	Percentage	Total Households
Housing Type	Overcrowding: Average house size	v22		Average	Total Households
Housing Type	Tenure: Owned or Shared Ownership		v31	Percentage	Total Households
Housing Type	Overcrowding: People per room	v23		Average	
Housing Type	Average number of rooms			Average	Total Households
Housing Type	Tenure: Rent (Private)	v17	v33	Percentage	Total Households
Housing Type	Tenure: Rent (Public)	v16	v32	Percentage	Total Households
Housing Type	House: Semi-detached		v26	Percentage	Total Households
Housing Type	House: Terraced	v18	v29	Percentage	Total Households
Socio-economic	Household: 2+ Cars	v26	v41	Percentage	Total Population (16 to 74)
Socio-economic	Economically Inactive: Full-Time Student	v31	v40	Percentage	Total Population (16 and over)
Socio-economic	Economically Active: Working full-time		v47	Percentage	Total Population (16 to 74)
Socio-economic	Economically Inactive: limited by long term illness	v29	v35	Percentage	Total Population
Socio-economic	Economically Active: Working part-time	v33	v46	Percentage	Total Population (16 to 74)
Socio-economic	Economically Inactive: Provide unpaid care	v30	v36	Percentage	Total Population
Socio-economic	Economically Inactive: Looking after family	v34		Percentage	Total Population (16 to 74)
Socio-economic	Economically Inactive: Unemployed	v32	v45	Percentage	Total Population (16 to 74)
Socio-economic	Occupation: Routine/Semi-Routine	v25		Percentage	Total Population (16 to 74)
Socio-economic	Travel to Work: foot, Bicycle or Other		v44	Percentage	Total Population (16 to 74)
Socio-economic	Travel to Work: Work from home	v28		Percentage	Total Population (16 to 74)
Socio-economic	Travel to Work: Private Transport		v43	Percentage	Total Population (16 to 74)
Socio-economic	Travel to Work: Public Transport	v27	v42	Percentage	Total Population (16 to 74)
Socio-economic	Qualification: Level 1, 2 or Apprenticeship		v37	Percentage	Total Population (16 and over)
Socio-economic	Qualification: Level 3		v38	Percentage	Total Population (16 and over)
Socio-economic	Qualification: Higher Education (L4)	v24	v39	Percentage	Total Population (16 and over)
Socio-economic	Employment: Agriculture/Fishing	v35	v48	Percentage	Total Employed Population
Employment	Employment: Administrative activities		v58	Percentage	Total Employed Population
Employment	Employment: Education		v60	Percentage	Total Employed Population
Employment	Employment: Financial intermediation	v40	v53	Percentage	Total Employed Population
Employment	Employment: Utilities		v55	Percentage	Total Employed Population
Employment	Employment: Hotel / Catering	v38	v51	Percentage	Total Employed Population
Employment	Employment: Health / Social work	v39	v52	Percentage	Total Employed Population
Employment	Employment: ICT activities		v57	Percentage	Total Employed Population
Employment	Employment: Manufacturing	v37	v50	Percentage	Total Employed Population
Employment	Employment: Mining/Quarrying/Construction	v36	v49	Percentage	Total Employed Population
Employment	Employment: Public administration and defence		v59	Percentage	Total Employed Population
Employment	Employment: Transport		v56	Percentage	Total Employed Population
Employment	Employment: Wholesale/retail trade	v41	v54	Percentage	Total Employed Population

Table 4.2: Potential candidate variables from OAC 2001 & 2011 and additional Energy domain variables relating to central heating fuel types, electricity and gas consumption and a fuel poverty indicator

possible designs are illustrated for each of the four stages. The variable selection process (Stage 2) involves complex data decisions and the investigation of correlation, distribution and geography. Uncertainties in the data relating to scale, geography and transformation are questioned and this introduces two new research questions to the thesis (RQ3 and RQ4). These complexities and sensitivities are investigated in the following chapter and the possibility of visualising them are investigated in the creation of a new theoretical framework and instantiation outlined in Chapter 6. An investigation of candidate variable sources reveals 78 variables useful for analysis at this time, with a number of other potential variables and sources likely to be available for future analysis.

5

Geography and Scale: Data Preparation

This chapter investigates the complexities and sensitivities associated with scale and geography within the variable selection process. Locally weighted statistics are explained and demonstrated in Section 5.1 and four stages of scale within the variable selection process are introduced in Section 5.2: input, standardise, locality and output. The use of multiple scales for the standardise stage and varying parameters for locality are discussed in Section 5.4 through the preparation of the 78 candidate variables, which were found to be appropriate for investigation for energy-based geodemographics in the previous chapter (Section 4.3). In Section 5.5 the four stages of scale are addressed in the context of the four stage process for the generation of geodemographics (as introduced in Section 2.3.4). The theoretical concepts and datasets described and prepared in this chapter are the foundations for the framework and interactive prototype described in the subsequent chapter.

5.1 Geographical Variation: Local Statistics

While global summary statistics are useful for a quick overview of the dataset, local statistics can be used to determine local patterns. In order to accurately represent geographical variation using local summary statistics the statistical calculation must take geographical location into account. Unlike global statistics, which use the whole dataset for calculating a value (such as an average or a correlation), local statistics take into

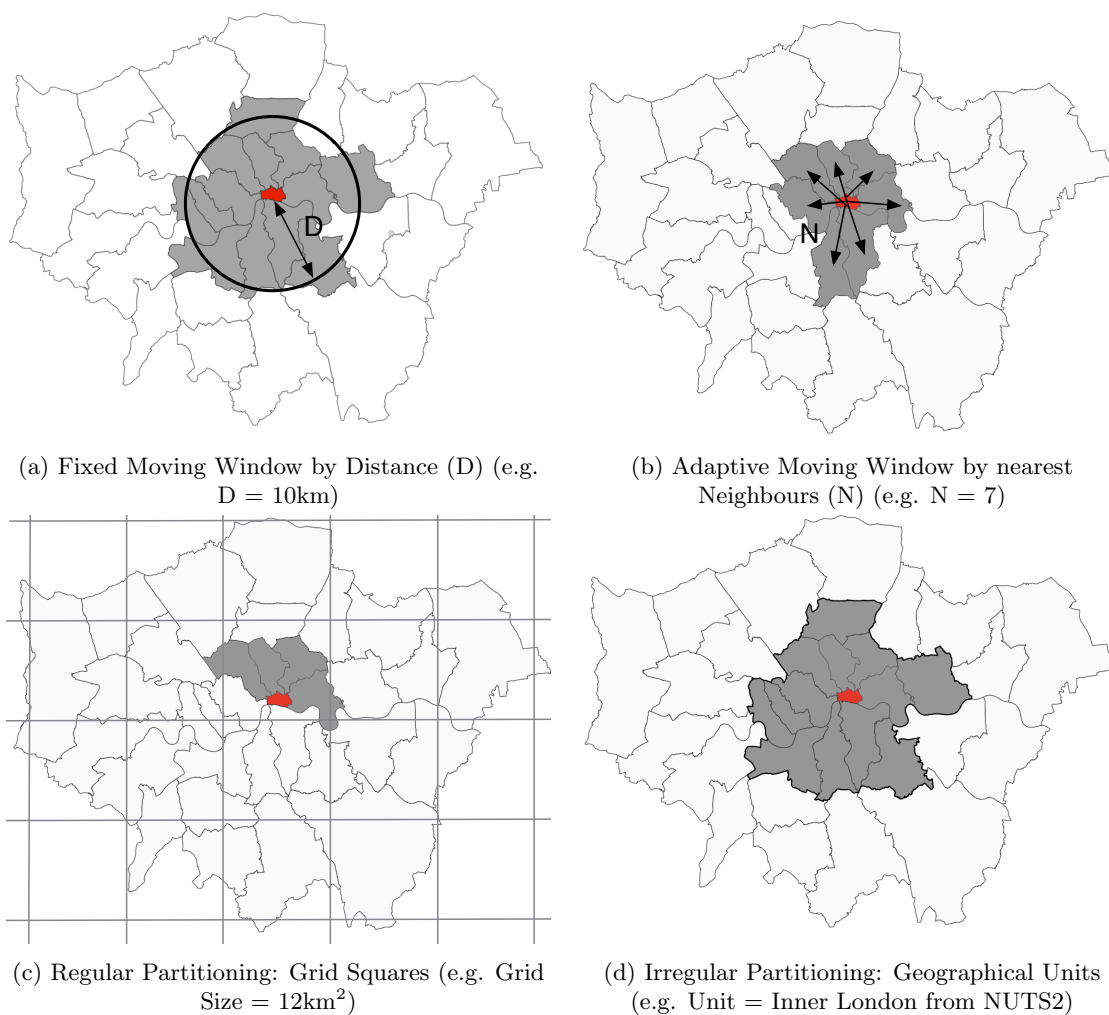


Figure 5.1: Four methods for local statistics calculation: Each example shows Local Boroughs of Greater London, the starting location of City of London (in red) and the locality region (dark grey). Locality changes depending on the chosen method and parameters (D, N or partitioning unit). Contains National Statistics data ©Crown copyright and database right 2015, OS data ©Crown copyright and database right 2015

account only a small number of data values within the vicinity of a defined area. The concept of this defined local area is termed ‘*locality*’ in this research.

5.1.1 Calculating Local Statistics

Four types of locality calculation have been identified from the literature (Fotheringham et al., 2002; Harris et al., 2010, 2011, 2014). These are shown as four equivalent illustrations on a small selected example in Fig. 5.1, where the region of City of London¹, is used as the started location, i.e. the location for which the local statistic is being calculated. The locality’s (shaded grey in Figs. 5.1a-5.1d) size and shape changes depending on the calculation technique, the parameters (value of D, N, or type and size of partitioning unit), as well as the method use to determine whether neighbouring units are inside or outside the defined vicinity. In the calculation for Fig. 5.1 (and subsequent calculations in Section 5.4.3) the population-weighted central location (point) of each geographical unit (polygon) is used to decipher whether the unit is inside or outside of the calculation area. In Fig. 5.1b the nearest neighbours are defined based on the shortest distance from the central point of the City of London and the central points of the neighbouring units. This definition could have used the absolute centroid point (not population-weighted) or by determining whether the polygon is entirely within, overlapping or majority-within the defined area. These point-in-polygon or polygon-in-polygon functions are standard GIS functions for determining whether geographical features are inside or outside of a given area. As the data in this thesis relates to where people live, the population-weighted centroids are used.

The two *moving window* approaches, as shown in Fig. 5.1a and Fig. 5.1b, calculate one output value (local summary statistic) for each geographical unit in the calculation. If the example shown in Fig. 5.1 was expanded to each of the 32 Boroughs of Greater London then 32 summary statistics will be produced. Fig. 5.1a uses a *fixed* distance measurement of value D to determine the regions within the defined area. The fixed distance ensures a consistent scale across the dataset. Fig. 5.1b takes into account the nearest N number of neighbouring regions and produces different sized locality regions. The adaptive approach ensures consistent sample size.

Fig. 5.1c and Fig. 5.1d represent a calculation which uses *partitioning*. Fig. 5.1c uses a regular grid square (or other type of grid could be used) approach as the calculation region. Fig. 5.1d uses a lower resolution geographical unit for the (irregular) partitioning area. The partitioning approach not only determines the size and shape of the locality region, but the number of output values. Expanding the example across the 32 Boroughs of Greater London results in an output value (summary statistic) for each of the 20 grid

¹Using 2011 Regions from ONS (ONS, 2011)

squares in Fig. 5.1c and for the two NUTS2 regions – Outer and Inner London – in Fig. 5.1d. Partitioning can be useful for reducing the quantity of output values as results can be based on aggregations of the original (input) data. Partitioning methods may also reduce data quality by introducing geographic error but sophisticated and accurate partitioning methods are possible (e.g Martin, 1989).

Coincidentally the neighbouring regions within the locality of Fig. 5.1d are identical to those in Fig. 5.1a. A slight change to the distance D or geographical unit of partitioning will result in a different locality size and shape in each case. This starts to demonstrate the sensitivity of the calculation based on the parameters and method used. In addition to the method used and parameter chosen (N , D or partitioning type and size), different types of weighting can be used in the calculation. Two types are discussed here: equal-weighting and distance-weighting. For equal, each region has identical weighting in the calculation. To reach more discriminating results, distance² can determine how much influence the neighbouring region has in the final calculation. Distance-weighting can be used in all four types of calculation. It is particularly effective in the moving window examples, where the input regions in the locality calculation are also the output regions. The inclusion of distance-weighting make the calculation more timely in terms of computer processing as the distance from each region to each other (i.e. a distance matrix) must be calculated, prior to the calculation of the local statistic.

5.1.2 Locality Calculation Sensitivity

In order to represent the sensitivity of varying the value of N in the calculation, a distance-weighted adaptive moving window approach is illustrated as an example. Fig. 5.2 displays the local correlation (Pearson's r) coefficient of the two DECC energy consumption variables: 'average annual gas consumption' and 'average annual electricity consumption' (variable details are described in Section 4.3.1). Firstly, the original LSOA³ data is aggregated to 326 Local Authority Districts (LAD) for all of England. Each of the 326 are included in the local statistic calculation. The global correlation for this pair of variables at the LAD level has a coefficient value of -0.32, yet the local statistics reveal that this negative correlation is not consistent throughout England. Fig. 5.2 illustrates how varying the number of nearest neighbours (N) from 100 to 50 to 25 reveals very differing local patterns.

When 100 neighbours are used in the locality definition (Fig. 5.2b), the regions around London appear to have a different correlation (positive rather than negative) from the rest of the country. For 50 neighbours (Fig. 5.2c) positive correlations in the northwest start

²Euclidean distance is used in the calculations in this thesis

³Lower Super Output Areas: The 2nd tier Census Geography, with areas of between 400-1200 households, aggregated from Output Areas (OA): <http://bit.ly/16XucgG> – as discussed in Section ??

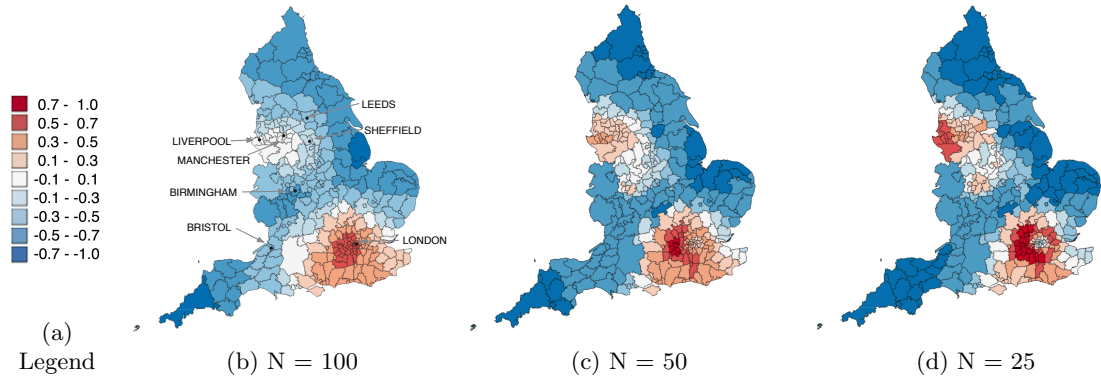


Figure 5.2: Local correlation coefficient of ‘gas consumption’ and ‘electricity consumption’ for 326 LAD in England using an adaptive moving window approach varying neighbours (N) from 100 to 50 to 25. Energy variables from (DECC, 2013a) Contains National Statistics data ©Crown copyright and database right 2015, and OS data ©Crown copyright and database right 2015

to appear, and for 25 (Fig. 5.2d) the difference between positive and negative correlations throughout the country intensify. In short, as the size of N is decreased in the locality calculation, more local patterns appear on the map. The example shown in Fig. 5.2 reveals increasingly positive correlations in densely populated areas (large cities are named in Fig. 5.2b) and increasingly negative correlations in more rural isolated locations. This strong spatial structure is particularly evident in this example and is to be expected, as there are lower levels of gas supply in isolated rural communities and to the apartment blocks, which dominate residential living in inner London. The example demonstrates that correlation can be highly geographically variant and that variation in phenomena – such as energy use – are detectable at different geographical scales.

The GWModel R package (Lu et al., 2014) was used for the calculations in this example as it incorporates distance-weighting in the calculation of local correlation coefficients (both Pearson’s and Spearman’s) using either a fixed or adaptive approach. The details of this package and the calculation process are explained in greater detail in Section 5.4.3. To illustrate the alternative calculation methods similar examples could have been created and compared for each of the other types of locality calculation (shown in Fig. 5.1), with differing values for D and size and type of partition. Equal and distance-weighting could also be compared. Each of these locality calculations will produce different outputs depending on the method, parameter and weighting chosen (Brunsdon, 1999). Although geographical variation combined with pair-wise correlation are important for geodemographic variable selection (see Section 2.4), exploring all of the alternative locality methods is beyond the scope of this thesis. For testing the inclusion of locality in this context the moving window distance-weighted approaches are used with variations of N and D , as discussed in Section 5.4.3.

The choice of N in the example in Fig. 5.2 was determined based on the total number of regions (326) within the dataset. 100 was chosen to represent the regional differences

in the data. The smaller subsets divided N by 2 each time to reach 50 and 25. The use of the adaptive approach ensures that the sample size remains 25 or more. A fixed distance approach may result in small and unreliable sample sizes in rural areas, as the statistical regions used in this example are based on population size. More sophisticated methods of parameter selection and testing for sample size instability is discussed as future work in Chapter 9. The fact that the data is aggregated to LAD level will also affect the results, with the original level of LSOA or an alternative geographical aggregation likely to produce different results. These varying factors relating to scale, as well as type of calculation and parameter value, are all important to consider in the process and flow of the data from input variables to output summaries ready for consideration for classification.

5.2 Dimensions of Scale

Aspects of scale are introduced and discussed in Section 2.3.3.2. For this research scale is separated into two types: *Scale Resolution* (SR) and *Scale Extent* (SE) (see Fig. 2.2). SR and SE can relate to attributes, time and space. SR refers to the degree of precision used to define the measurement of the data, which is determined by the interval of sampling or imposed by aggregation, e.g. combining time-based data into days of the week, aggregating variables into grid cells, or grouping age-group attributes. SE refers to the scope of analytical focus, e.g. the geographical extent, the total length of time period, or breadth of categorical information. For energy-based geodemographic variable selection spatial scale is important.

5.3 Spatial Scale

Spatial SR and SE of the data variables used in the classification will affect the ability to accurately portray the geographical variation of consumers. Varying scale through aggregating or filtering the dimensions of the data (see Section 2.3.3.2 and Fig. 2.2) will also impact the summary statistics. The smaller the resolution (SR) of geographical unit, the more detailed the dataset and increased likelihood of a greater variance in data values. Upon aggregation the data outliers are removed and the statistics become generalised. Changing the aggregation of the data can produce the MAUP effect (as discussed in Section 2.3.3.2). Choosing the optimal level of aggregation for analysis is important, as the scale must reflect the analytical extent and geographical detail necessary for the use case, as well as the scale of variation of the phenomena being analysed. The scale used for investigating UK-wide patterns in data is different from that needed for local analysis. The OAC 2001 ‘Multicultural’ Super Group is one of seven profiles representing a proportion of the UK population. If the SE is reduced to London, the ‘Multicultural’ group is heavily over-represented, while the ‘Countryside’ group heavily under-represented. If the same 41

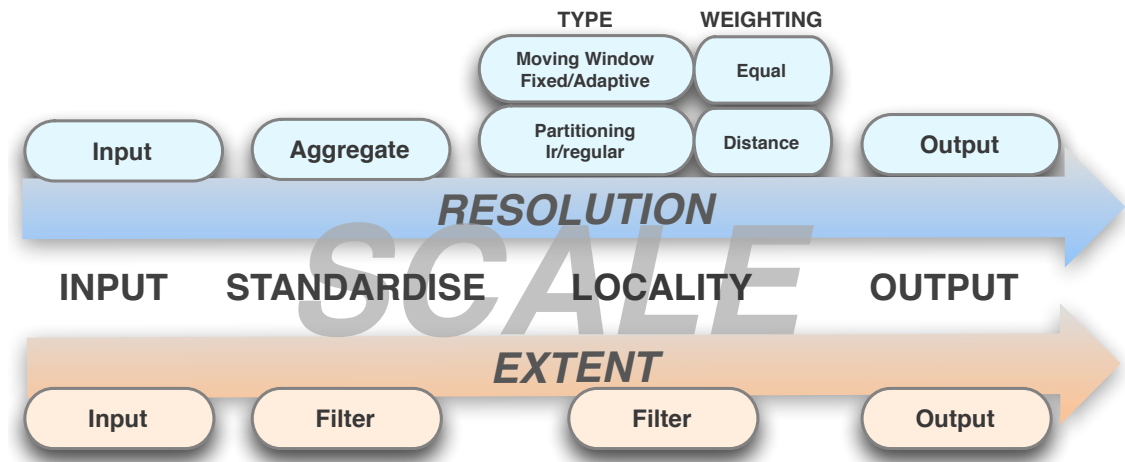


Figure 5.3: Four stages of the process: Input, Standardise, Locality and Output, each with two dimensions of Scale: Resolution and Extent

OAC 2001 variables are used to create the same number of clusters only for London, the resulting profiles are very different, as shown in the creation of a London specific OAC (LOAC) (Petersen et al., 2011). This demonstrates that varying the SE of a dataset during the classification process can have great effect on the output.

The visualisation of the data at different scales will allow users to understand the impact that scale can have on single data variables, as well as their pair-wise or multivariate relationship. SR and SE are investigated in the geodemographic variable selection process in the following section.

5.3.1 Scale in the Variable Selection Process

In addition to utilising local as well as global statistics to incorporate geography, the SR and SE of the data are important to consider across the process from input to output. There are many instances where the SR and SE of the data can be varied. As this is relatively arbitrary, depending on the data and geographical units of analysis, the only approach is to show different solutions and look for variability and stability. Hence the necessity for specifically designed visualisation. The flow of the process is illustrated in Fig. 5.3, where four stages of the variable selection preparation process (Input, Standardise, Locality and Output), each contains two dimensions of scale (SR and SE).

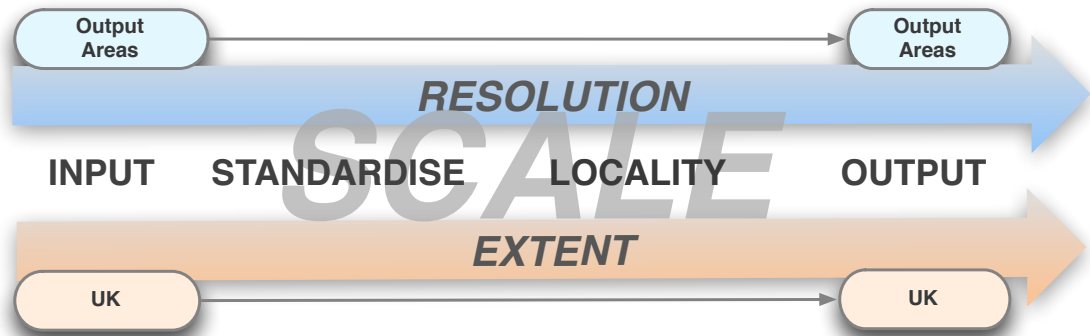
- **Input:** resolution (IR) and extent (IE) refer to the smallest areal unit and full extent of the raw (source) data. The Census 2011 variables used in this analysis are available at the IR of OA for the IE of England and Wales. The DECC variables are available at the IR of LSOA and the consumption variables cover an IE of England and Wales, while the fuel poverty variable only England.

- **Standardise:** resolution (StR) and extent (StE) refer to the chosen standardised scale of the analysis. As examples, variables can be aggregated to a StR of LAD as shown in Fig 5.2, or reduced to the StE of London as described for the LOAC example (Petersen et al., 2011).
- **Locality:** resolution (LR) and extent (LE) allow for the calculation of summary statistics at varying local as well as global scales. In Fig 5.2, the LR adapts from 100 to 50 to 25 neighbours and the LE remains the same as the StE (England). If a LE is defined at this stage the local statistics are only calculated for the given extent, i.e. local statistics could be calculated only for London as additional discriminating factors are needed for London compared to elsewhere in the UK. Although this option would be useful for some applications it adds additional complexity. Varying LE is not investigated in the analysis in this thesis..
- **Output:** resolution (OR) and extent (OE) refer to the dimensions of the data after it has been through the previous stages and is ready for analysis. If no changes to scale are made during later stages then OR is equal to IR and OE to IE (see Fig. 5.4a). If data is aggregated (or filtered) at the standardise stage then OR is equal to StR (see Fig. 5.4b). Utilising the locality stage only effects the output scale if the resolution of the partitioning reduces the number of output items, therefore OR is equal to the partitioning size IR. The example from the literature in Fig. 5.4c uses moving window and therefore OR remains the same as StR.

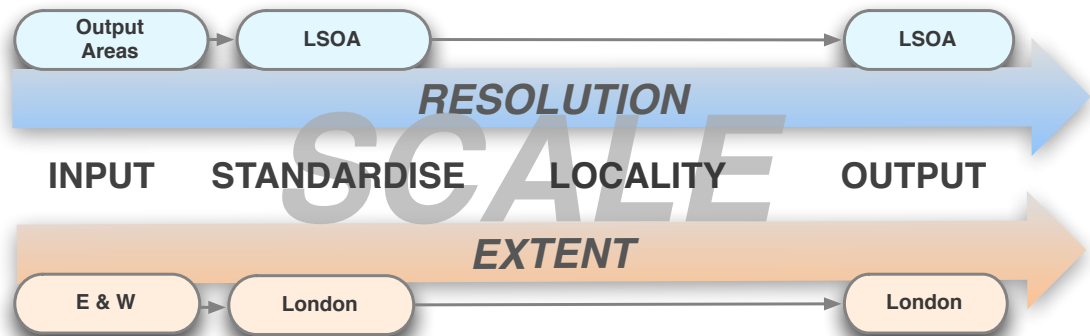
5.3.2 Scale Sensitivity in Variable Selection

Adjusting the scale through the aggregation and/or filtering of data based on geography, time or attribute at the Standardise and Locality stages allows the associated sensitivities of scale to be explored during the geodemographic variable selection process. Although spatial scale is the main focus in this analysis, both time and attribute are important. Temporal aspects could not be investigated in detail for the analysis as detailed time-based data is not available in the Census variables or the DECC energy variables considered (see Section 4.3). Yet, temporal scale must be considered for future datasets, as detailed consumption data will allow the annual average measurements from DECC to be combined with seasonal, daily or hourly measurements of each appliance.

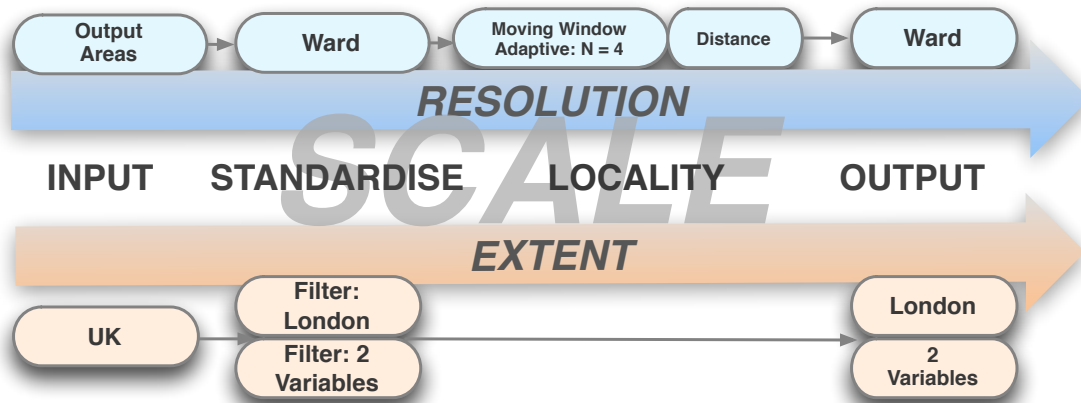
The consideration of attribute aggregation and filtering is also important. Combining or separating variable groups to reduce strong correlations may make more discriminating profiles (relating to US#46 and US#47 from Section 4.2.2). This is demonstrated in the variable changes for OAC 2011, where separate ethnicity variables are used instead of combined as in OAC 2001 (Gale, 2014b). Attribute-based aggregation and filtering would



(a) Output Area Classification (OAC) 2001 and 2011 (Vickers et al. 2005, 2007; Gale, 2014)



(b) London Output Area Classification (LOAC) 2001 (Petersen et al., 2011)



(c) Example of Spatially Weighted Geodemographics for London (Adnan et al., 2013)

Figure 5.4: Four stages of the process for three differing geodemographics referenced in the literature

also be useful for understanding the vast amount of data from smart homes, as shown with the need to group household profiles by type as well as time in the Smart Home HeatLines prototype (see Section 3.5.2.1) and the use of different grouping options in Consumption Signatures (see Section 3.5.2.2).

Altering spatial-, temporal- and/or attribute-based scale may affect the variables selected for geodemographic classification. The degree to which results change based on aggregation or filtering (affecting SR and SE) has had limited investigation in academic research and the inclusion of geography by utilising local statistics has yet to be fully investigated for geodemographic variables. Therefore, varying spatial scale is considered in detail for the framework described in the following chapter.

5.3.3 Spatial Scale in Geodemographic Examples

In order to demonstrate how these aspects of scale are linked to the current open geodemographic examples in the literature, the scale diagram (Fig. 5.3) has been modified to illustrate three examples (Fig. 5.4). The first (Fig. 5.4a) represents OAC 2001 and 2011 which uses the standardised extent of the UK ($IE = OE$) and the IR of OA as the OR (Vickers et al., 2005; Vickers and Rees, 2007; Gale, 2014b). The second (Fig. 5.4b) filters the IE to London at the SE stage, but retains the IR of OA to create a London specific OAC (known as LOAC) (Petersen et al., 2011). Lastly Fig. 5.4c illustrates an experiment for geographically-weighted geodemographics, where the locality stage is utilised. The standardise stage includes spatial aggregation (StR) and two types of filtering (StE). The OA data is aggregated to LAD. Attribute extent is filtered from 41 to 2 variables, with spatial filtering used to reduce the geographical extent to London. A moving window distance-weighting approach with 4 nearest neighbours (Adnan et al., 2013) is used to calculate local statistics. As a moving window approach is used the OR is equal to the StR as there is an output value for each region in the calculation.

5.4 Data Preparation

The 78 candidate variables described in Section 4.3 (Table 4.2) are used in this section to demonstrate the geodemographic variable preparation process for the four stages of scale: input, standardise, locality and output. The variable list from Section 4.3 is available in Appendix B.9 in the default order used for the preparation, with additional source information, and amended domain and variable names. The ‘employment’ and ‘socio-economic’ domains were combined for this analysis, as types of employment are less likely to influence energy consumption than housing type, tenure, composition and age of the population. Additional sub-domains are also created through the naming of the variables, such as employment, age or ethnicity.

5.4.1 Input Scale

At the time of this analysis the 75 Census variables were available at OA 2011 level for England and Wales (Scotland and Northern Ireland data has since been released). All three DECC variables were available at LSOA 2001 level⁴, the ‘fuel poverty’ variable only for England, and the consumption variables for England and Wales. The fact that the two data sources (DECC and Census 2011) use UK Census boundaries from different Census years results in uncertainties when merging the data for geographic comparison. Whilst the 2011 OA boundaries were kept as close to 2001 OAs as possible, some regions were amended to reflect administration changes, population growth and statistical homogeneity (discussed in Section 2.3.3.2). When referring to LSOAs, of the 348 local authorities in England and Wales, 54 have amendments to 5% or more of their LSOAs, while only 6 LADs have no edits at all (Tait, 2012a). The consideration of how to merge the data accurately is important to ensure the reliability of the data being compared.

2011 Census boundaries are used for this analysis as they are the most recent and therefore most realistic to represent the current population. The Census variables are also more precise and reliable than the Energy (consumption and fuel poverty) data from DECC, which is termed as ‘experimental’ at the LSOA level (DECC, 2013a,c). The estimation of 2001 LSOA level to the 2011 geographies cannot be truly geographically accurate due to known issues associated with data aggregation (as explained in Section 2.3.3.2). Yet as proof of concept for analysing multiple variables for energy-based geodemographics the energy data is converted to the 2011 boundaries. This method chosen attempts to keep errors to a minimum. One fundamental issue with changing boundaries is that many of the areas are split or shifted and therefore exact matching of LSOA 2001 to LSOA 2011 is not possible. For this reason, it is more appropriate to allocate the data to the smaller 2011 OA boundaries and then re-aggregate these to the larger 2011 LSOA. This also enables the added benefit for the comparison of all the variables at OA level, despite the energy data being disaggregated to at this resolution.

The energy data was allocated to the 2011 OAs using QGIS⁵. The OA 2011 boundaries were overlaid on the LSOA 2001 boundaries and population-weighted OA centroids used to determine which OAs are located within each LSOA, using a point-in-polygon query. The consumption and fuel poverty data was then allocated to the OAs by allocating an equal proportion of the data to each household, using the Census 2011 household count. Once the DECC variables were estimated to OA 2011, all the 78 variables were combined into one table, along with the five denominators (see Appendix B.9).

⁴DECC variables are now available at 2011 boundaries.

⁵A Free and Open Source Geographic Information System available at <http://www.qgis.org/en/site/>

	European Regions Level 1 (NUTS1)	European Regions Level 2 (NUTS2)	Local Authority District (LAD)	Lower Super Output Areas (LSOA)	Output Areas (OA)
Number of Regions in England	9	30	326	32,844	171,372

Table 5.1: The five aggregations of SR used in the standardise stage (StR)

5.4.2 Standardised Scale

Alternative levels of scale are included at this stage in order to investigate whether varying the scale of the data can affect variable selection. Firstly, a decision was necessary for the StE. The fact that the ‘fuel poverty’ variable was not available for Wales meant that either the variable was included as only a proportion of the full extent, the variable was removed, or all other variables were reduced. The investigation of varying SE was deemed to be less important in this case study than the investigation of varying SR, as SE in geodemographics has already been investigated in some detail by those investigating local-based geodemographics, such as LOAC (Petersen et al., 2011). Therefore, the full extent of England was used for the investigation at this stage following the OAC principle that variables should all be at 100% of the extent (Vickers et al., 2005). This leaves an open extension to filter the data and test the comparison of varying StE and LE in the future, such as comparing results for an StE of London to those of Manchester.

In terms of SR, five hierarchical geographical aggregations of OA, LSOA, LAD, NUTS2 and NUTS1 (see Table 5.1) were chosen to allow the sensitivities of varying SR to be investigated. OA and LSOA were chosen as they reflect the SR of the input data. Since the release of the 2001 UK Census at these levels, these two geographies have become popular for detailed neighbourhood level analysis in the UK. The third tier geography of MSOA was not chosen for this investigation as it is less commonly used in case studies and it is not one of the official geographies for the release of ONS area classifications (as described in Section 2.3.3.2). LAD is one of the official ONS area classifications and therefore was chosen for the next aggregation of StR, with a total of 326 LADs in England. Health Areas, the final official ONS area classification geography, do not fit into the hierarchical structure of OA, LSOA and LAD and therefore are not used in this investigation. Two higher regions of NUTS2 and NUTS1 were chosen as they fit within the geographical hierarchy in the analysis (OA, LSOA and LAD) and illustrate much larger aggregations of the variables for regional-based analysis. Although, tested in the process, NUTS1 was eventually dropped as the 9 regions were not enough for the local statistics to be calculated (as explained in Section 5.4.3) and both the global and local correlations are likely to be unreliable in the comparison: *“for correlations computed from variables which both have fewer than 10 scaling points, the amount of information loss can be substantial”* (Martin,

1978, pp.307). An extension to this research could involve the representation of non-hierarchical geographical units, commonly used in government for reporting and decision-making, as irregular geographic partitioning into entities complicates multivariate analysis and is likely to heavily influence the results.

5.4.3 Locality Scale

There is a vast array of locality combinations possible to explore as described in Section 5.1. The two moving window options produce an output value for each of the regions (defined in StR), which can lead to more spatially precise results than a standard partitioning option (although sophisticated options are available, e.g. Martin, 1989). Partitioning can also introduce an additional scale into the analysis (LR), which may affect the sensitivity analysis results. The moving window options were chosen for this investigation in order to keep the LR consistent with the StR. In order to demonstrate the concept in this context, the moving window approaches with a few variations of N and D are tested.

The two types (fixed or adaptive) are incorporated in the GWModel R package (Lu et al., 2014), which the calculation of geographically-weighted local correlation coefficients (both Pearson's and Spearman's). The gwss (geographically weighted summary statistics) function in GWModel builds an inverse distance matrix for all the locations in the dataset and then calculates the statistics based on either a given number of neighbours or a fixed distance (Lu et al., 2014). For this calculation the population-weighted centroid points were used instead of the whole polygon in order to speed up the calculation process, yet was found to be problematic for the two larger datasets due to the sheer volume of connections.

Due to the large number of geographical units in LSOAs and OAs (see Table 5.1), the calculation process was found to be process and time intensive. The full process was only completed for NUTS2 and LAD level. When building and testing, the R code for the adaptive and fixed methods for NUTS1 regions was problematic as the centroids are so far apart and there are only 9 regions. The NUTS2, with 30 regions, was therefore used to run and test the algorithm. Initially the program was written and tested on few variables and eventually run for all 78. The GWModel package and functions are designed for creating geographically-weighted statistics for the comparison of two variables. This was run for all 78 variables and the Pearson's correlation coefficient extracted. As the distribution of the variable is important to the process in addition to the correlation coefficient, the same methodology was repeated for the analysis of local distributions by calculating the skewness value at each stage. Skewness was chosen as the statistic to analyse distribution as the GWModel package had only this option for a measure of distribution included in the gwss function.

In each case, the inverse distance matrix for all regions must be first created. The initial code for NUTS2 took approximately 30 minutes using a new MacBook Pro with 8GB RAM. This was then used on the LAD dataset with 326 regions, which initially took over 12 hours to run the calculations (tested overnight). Despite some optimisation improvements to the code, the next step to LSOA with over 32,000 points was too large for the resources and time available. The code stalled at building the inverse distance matrix for all regions and was therefore terminated. Further improvements are possible including re-writing the function to calculate the distance calculation in a more efficient manner as well as testing the code on multiple cores or a faster processor with more RAM. Rather than spending more time trying to calculate data for LSOA or lower it was seen as more beneficial to demonstrate the framework using the local data for the two levels of LAD and NUTS2 created. The limitations of the methodology and datasets are discussed in more detail in Chapter 9.

Three sets of varying neighbours ($N = 5, 10$ and 15 for NUTS2 and $N = 25, 50$ and 100 for LAD) and two varying distances (100km and 200km) were calculated for each type. It is noted that the local correlations for NUTS2, particularly when $N = 5$, may be subject to substantial information loss due to the low numbers of calculation points (Martin, 1978). Improvements can be made to the calculation by using alternative correlation methods which correct for scale (Martin, 1978); however, these are not available in the GWModel package and were therefore not produced for this analysis. Sample size is nonetheless taken into account when investigating sensitivity in Section 7.3.2. The other three aggregations (NUTS1, LSOA and OA) have no local summaries and therefore do not allow for advanced analysis of geographical sensitivity, yet they remain useful to the analysis of the energy variables and the sensitivity of scale. The global skewness value and Pearson's correlation coefficient were calculated for each of the five SR.

5.4.4 Output Scale

As a summary of the variation of scales in the prepared data, Fig. 5.5 illustrates each stage of the process. These output datasets are utilised in the final prototype (Chapter 6) to investigate the sensitivities in the variable selection process, in order to answer RQ4.

5.5 Data Calculation Method

The simplification of the geodemographic data preparation process to the four levels of scale (in Fig. 5.5) does not allow the whole process of the calculations to be fully explained. For the data preparation process explained in the previous section, a step-by-step diagram has been produced (Fig. 5.6) with reference to the decisions based on scale previously mentioned in Section 5.4. This method is described in the following section.

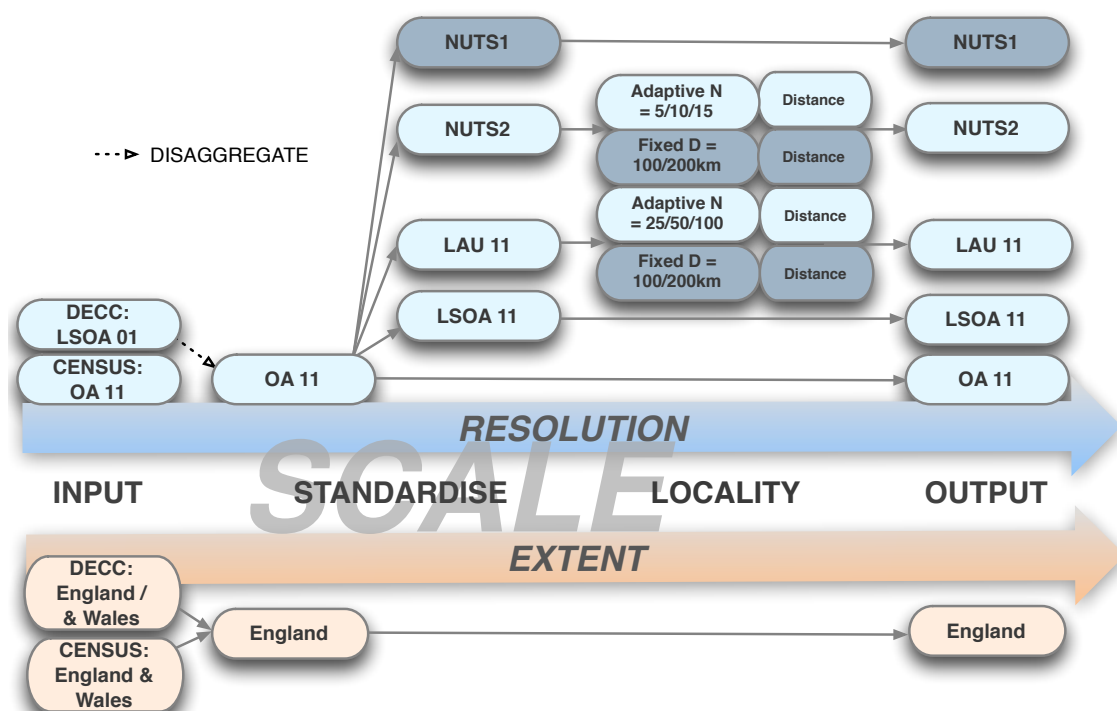


Figure 5.5: The data scale resolution and locality options used in the four stages of data preparation. The darker blue options were not included in the final prototype.

5.5.1 Describing the Method

Fig. 5.6 illustrates the method for the creation of the data using R software. First the 75 census variables were calculated (it was necessary to combine some of the Census variables to create the exact OAC variables) and merged from their separate tables (see source tables in Appendix B.9) along with the 5 denominators. Then, the 3 variables from DECC were merged and converted to 2011 OA regions (as explained in Section 5.4.1). These 83 variables were combined into one table and the OAs for Wales were removed. The variables in the table were aggregated four times to produce 5 outputs – OA, LSOA, LAD, NUTS2 and NUTS1 – and then each variable was converted to a percentage based on their denominator (see Table 4.2). At this stage the data was saved as filenames ‘RAW’ with the geography abbreviation attached (OA, LSOA, LAD, NUTS2 and NUTS1). These tables were then duplicated; one was converted to the range standardisation in order for all variables (including ratios) to be comparable, while the second was initially transformed using the log algorithm and then standardised using the range standardisation following the OAC 2001 methodology (Vickers et al., 2005). The two datasets allow for the comparison of data transformation in the prototype. Transformation is important to understand when choosing variables for the geodemographic process as discussed in Section 2.3.3.3. As three types of transformation were tested as part of the amendment of the methodology for OAC 2011 (as described in Section 2.3.3.3, Gale, 2014b) it was

seen as less important than representing variations in scale and geography for this thesis. Therefore only the logarithmic scale was tested in comparison with non-transformed data (explained further in Section 6.1). These tables were saved as filenames ‘RNG’ and ‘LOG’ with the geography abbreviation, ready for use in the prototype (as shown in Fig. 5.6).

In summary, there are three output files for the 78 variables for each of the SR for the single SE of England. The RAW files are kept for reference. Two comparative datasets are available with RNG and LOG containing all 78 variables and the LOG variables having been transformed to the logarithmic scale. At this stage, each of the output tables for the non-transformed (RNG) and the logged (LOG) datasets are used for global calculations, while only two of the five have local calculations. The 326 LAD and 30 NUTS2 regions (demonstrating $L = \text{Micro}$ and $L = \text{Macro}$, which becomes clear when the framework is introduced in the following chapter) are used to calculate local summaries for both skewness and correlation for each region for each variable pair for each variation of N (as described in Section 5.4.3). Further derived statistics are created for each of the global and local outputs, as discussed in the following section.

5.5.2 Statistical and Structural Outputs

In addition to the global and local skewness values and Pearson’s correlation coefficients described in Section 5.4.3, further global descriptive statistics were calculated for each SR for both the RNG and LOG outputs, to allow for quick comparison. Descriptive statistics were calculated using the R package fBasics, which includes quartile ranges, mean, median, variance, standard deviation, skewness and kurtosis values. These values were stored for each SR and used as a look-up table to describe the global characteristics of the variable.

For the comparison of pairs of variables only the correlation coefficient was calculated, but descriptive statistics (as above) were calculated for these correlation coefficients for the global and local datasets and stored for comparative purposes. In terms of structure, the correlation coefficient calculation compares each of the 78 variables to each other, which is the equivalent of 6084 combinations (including self-to-self and duplicates as the correlation of Var1 and Var2 is the same as Var2 and Var1). Even after removing all duplicates and self-to-self correlation, the most efficient storage of the global correlation statistic is a matrix of 78 rows by 78 columns, where the diagonal is always 1.0 (as the correlation of a variable to itself is always perfectly correlated). The local correlations are more complex. Here, the most efficient storage is a table with a column for each variable to variable combination, removing duplications and self-to-self, with 30 and 326 rows for the NUTS2 and LAD datasets respectively.

In addition to the global and local skewness value and correlation coefficient, which refer to the distribution and correlation of variables, a global and potentially local

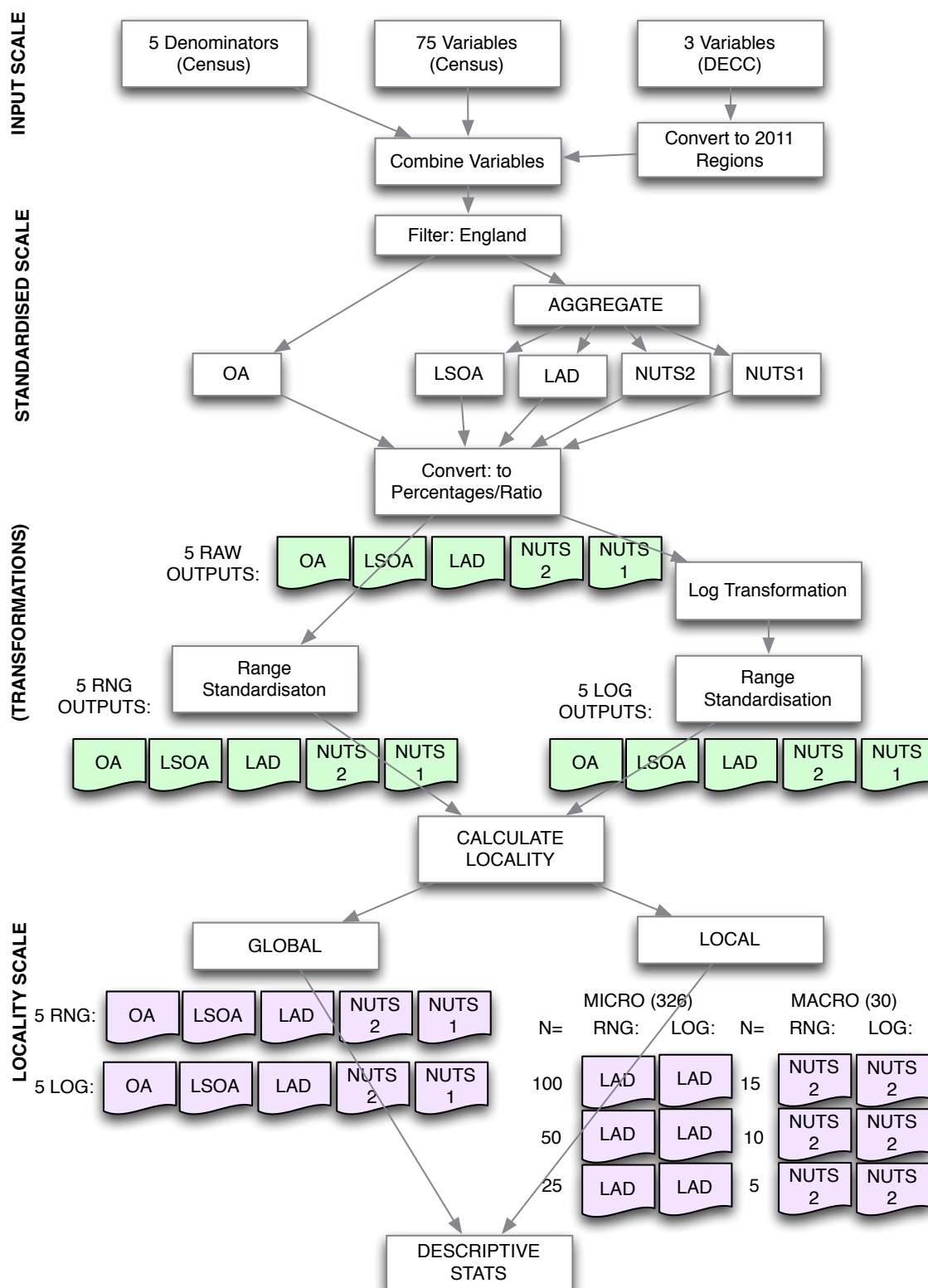


Figure 5.6: A flow diagram illustrating the data preparation process including the decisions for scale, locality and transformation

statistic to represent the geographical variation of the dataset was investigated. Options for spatial auto-correlation are explored and discussed in Section 2.4.3. Moran's I statistic was used to calculate a global value for each variable using the R Package *ape* (Paradis, 2014). Moran's I was chosen for this prototype as the value has a standard range from -1 (anti-correlated) to +1 (fully correlated), it is not as sensitive as other algorithms at the local level (e.g. Geary's C) (Anselin et al., 2002) and the global and local statistics have been applied in visualisation examples. As in the *gwss* function (explained in Section 5.4.3), an inverse distance matrix calculation is created in order to calculate the distance from each location and therefore calculate the spatial auto-correlation. The same constraint of time and resources (as discussed in Section 5.4.3) applies to the calculation of Moran's I as the inverse distance matrix is needed in the calculation. Therefore Moran's I was calculated only for NUTS2 and LAD, as these also have local correlation and skewness values. To calculate the global Moran's I, and the local statistics discussed in Section 5.4.3, for LSOA and OA, the SE of the region needs to be reduced or a different LE chosen, e.g. partitioning to a larger grid size. These are not tested in this analysis due to time and resource constraints, but the trade off between processing power and visualisation power is discussed in more detail in Section 9.3. The local Moran's I statistic was not explored for the prototype at this stage as local statistics (correlation and skewness) which show geographical differences in the variable were already produced. The use of local spatial auto-correlation (using Getis-Ord) for creating geographically weighted geodemographics has also already been investigated in the literature (Adnan et al., 2013). Further exploration of global and local Moran's I (or equivalent spatial auto-correlation calculation) would be useful in the context and remains an extension for future research.

5.5.3 Four Types of Scale in the Four Stage Process

Fig. 5.7 (black outline) illustrates how the four stages of scale, discussed in this chapter, fit within the four stages of the process of generating a geodemographic classification, shown in Fig. 2.3. Fig. 5.7 shows a detailed flow of the four stage process from Stage 1: the data preparation stage to Stage 3: the clustering stage. The stages outlined in Fig. 5.7 are all options prior to clustering and the sensitivities referring to scale, transformation or geography associated with each activity are displayed. The initial data pre-processing stage is run on the input data (IR/IE) and then the data is aggregated and/or filtered to the standardised level (StR/StE). At this stage the data is converted to percentages, indexes or equivalent to ensure the data is comparable (e.g. percentages in OAC 2001 and 2011). Variables are then transformed and/or standardised to the same scale (e.g. log transformation and range standardisation in OAC 2001 or inverse

hyperbolic sine and range standardisation for OAC 2011). The type of locality is chosen once the data is scaled and transformed. Then the global and local statistics are calculated to allow for the variables to be compared (at the output scale). Multivariate comparison is necessary to select appropriate and identify problematic variables for the classification. Once variables are chosen then clustering can commence (and there are many additional options to consider at the clustering stage as discussed in Section 2.3.3.4). The arrows prompt the creator to step back and alter the decisions from the standardised, data transformation or locality stages, all of which can change the scale of output and potentially the variables selected for clustering.

5.6 Chapter Summary

This chapter has investigated the concept of locally weighted statistics to identify geographical variations in the datasets and explore sensitivities associated with the definitions of locality, as well as the investigation of scale through both SR and SE at four stages of the classification generation process. These concepts are described and explained in the context of geodemographics. The defined area of a locality is influenced by the type of calculation, the number of N , the weighting as well as the SR and SE of the data. Varying each can change the result of the output. These concepts are all demonstrated through the preparation of the data variables for use in the prototype to represent the framework, as described in the following chapter.

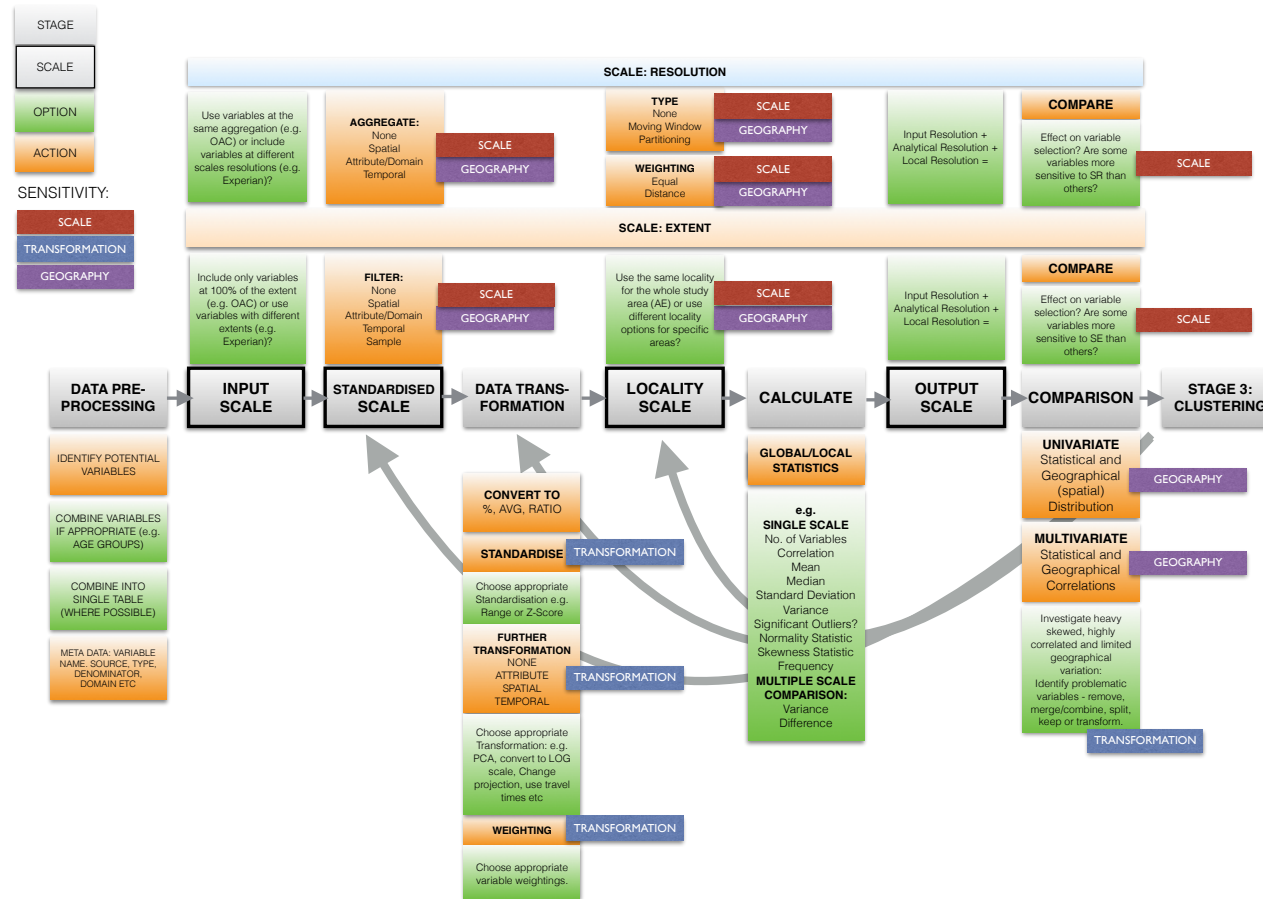


Figure 5.7: The four types of scale can be seen (with a black outline) in the flow of the geodemographic process prior to clustering, i.e. from Stage 1 through to Stage 3. Three types of sensitivity – scale, geography and transformation – are highlighted.

6

Building and Visualising the Framework

This chapter outlines a framework for visualising multivariate geographical data with particular attention to comparing local geographical variations of data, as well as allowing for the consideration of multiple scales and transformations. The framework is general and acts as a guideline for future multivariate geographical analysis. To demonstrate the framework, it is instantiated through a visualisation prototype in the context of variable selection for energy-geodemographics, using the data prepared in the previous chapter. A number of key elements discussed in the previous chapter are investigated for the framework including the calculation of global and local statistics, the number of variables to compare versus screen space, the types of scale and types of data transformation.

In order for the human eye to visually compare multiple variables, especially variables with multiple dimensions and many data items, the use of data reduction is fundamental. This allows for better comprehension of the sheer volume of data and enables the data to be visualised. The use of data aggregation and filtering are critical to representing the framework, whether this is by geography, time or attribute. As the framework is designed in the context of variable selection for the generation of geodemographics, the nature of the data means that geographical (spatial) aggregation and filtering are the main focus for the examples represented in the design of the prototype. Extensions for data with temporal- or attribute-based scale are discussed. The framework is flexible to other

applications of multivariate comparison where geography is a consideration. Additional scenarios applicable to the framework are discussed in Chapter 8.

6.1 Building the Framework

The parameter space associated with multivariate geographical comparison and elements of scale discussed in previous chapters is shown to be complex. The dimensions are combined here with knowledge of data visualisation to create a framework for visualising multivariate data across multiple scales with the inclusion of local geography.

Firstly, the representation of multiple variables in combination with their local values is difficult to comprehend and therefore it is simplified by considering the dimension of locality in three broad and loosely delimited bands: ‘*Global*’ (as used in cases where local variations are not considered), ‘*Macro*’ and ‘*Micro*’. ‘*L*’ represents the number of locations at which statistics are calculated (i.e. the locality region). When global values are represented $L = 1$, when local values are represented either $L = \text{Macro}$ (more than 1 but less than many locations) or $L = \text{Micro}$ (many locations). The threshold, at which Macro becomes Micro is flexible and depends upon a number of constraints including the number of variables being shown (V), the number of data points in the comparison, the visualisation represented, the user’s experience and monitor/display resolution possibilities. These are all typical constraints when designing data visualisation as discussed in Section 2.2.

In combination with the three bands of L , the number of variables (‘ V ’) are also grouped into four types: ‘*Uni*’, ‘*Bi*’, ‘*Multi*’ and ‘*Many*’. Uni is necessary as this represents uni-variate analysis and global or local calculated values representing the distribution of a single variable, e.g. the mean, median or skewness of a variable. Bi relates to the standard bi-variate comparison for instance in this case the correlation of two variables. Multi refers to multivariate analysis (explained in Section 2.4) where multiple variables are compared and the visual representations must change for instance from a scatterplot for bi-variate comparison to a SPLOM (see Section 2.4.2). When V becomes too many (hence $V = \text{Many}$) the representation must adapt to better use the screen space and allow for the comparison.

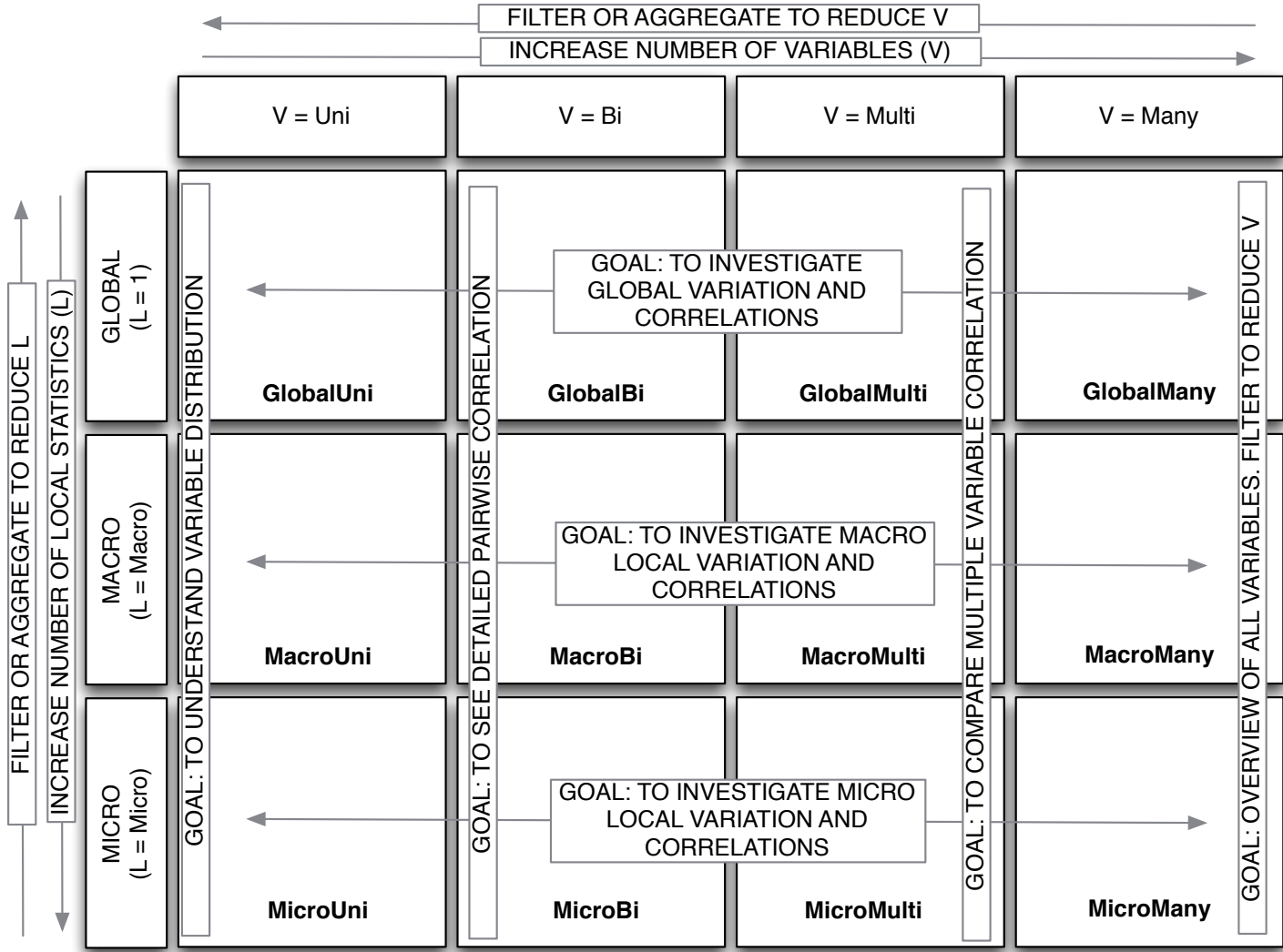


Table 6.1: The layout of the framework with the names of the cells and the goals of the three rows representing number of locations where local statistics are calculated (L) and the four columns representing numbers of variables (V)

6.2. POTENTIAL VISUAL REPRESENTATIONS OF THE FRAMEWORK

The proposed framework is illustrated as a table with L represented in three rows (Global, Macro and Micro) and V represented as four columns (Uni, Bi, Multi and Many), as shown in Table 6.1. The goals for each column/row combination are also summarised in Table 6.1 and the cells of the table are named (and referred to during the thesis) with the top row as *GlobalUni*, *GlobalBi*, *GlobalMulti* and *GlobalMany*, the middle row as *MacroUni*, *MacroBi*, *MacroMulti* and *MacroMany* and finally the last row as *MicroUni*, *MicroBi*, *MicroMulti* and *MicroMany* (see Table 6.1).

The ability to make comparisons when crossing the parameter space (shown in Table. 6.1) and the possibilities of transitioning from one part of the framework to another is explored through the investigation of data scale (both SR and SE) and transformation (T). This is represented in Table 6.2 where the ability to make visual comparisons of datasets with multiple scales (resolution or extents) or multiple transformations is expressed using a textual scale of ‘*Many, Some, Limited and None*’. Meaning that ‘many’ scales or transformations can be represented and compared by the viewer through to ‘none’ – where the visual comparison is near-impossible given the number of values. Table 6.2 shows that it is more difficult to compare data scale or transformation as the number of variables (V) or the number of local summary statistics (L) increases. The textual scale is also symmetrical diagonally across the table. This ability to explore the data must therefore be reflected by the adaption of the visual representation at appropriate thresholds. The possibility for visually encoding this parameter space is vast and potential designs are explored in the following section.

6.2 Potential Visual Representations of the Framework

In this section visual representation possibilities are discussed within the context of variable selection for geodemographics. Given the need to compare geographical variation with global and local statistics, two types of visual representation are proposed to visualise the framework: statistical and spatial. In order to transition through the framework and increase the number of L and/or V there are a number of data reduction possibilities including spatial or statistical aggregation, sampling or filtering by attribute, geography or time, as well as re-projection options such as using a cartogram instead of a choropleth for better use of space (as shown in Section 2.4.3). Possible visual representations for both statistical (top row) and spatial (bottom row) views are proposed in Table 6.3 which includes the use of matrices (Melville et al., 2011), scatterplots (Monmonier, 1989) and maps (Dykes and Brunsdon, 2007).

Standard well-known visuals are used in the proposal despite many other potentially more innovative alternatives being possible. These representations are proposed in Table 6.3 after investigating the literature (in Section 2.4) and reflecting on the feedback

		<div> <div>← FILTER OR AGGREGATE TO REDUCE V</div> <div>INCREASE NUMBER OF VARIABLES (V) →</div> </div>			
		V = Uni	V = Bi	V = Multi	V = Many
<div> <div>FILTER OR AGGREGATE TO REDUCE L</div> <div>↑ INCREASE NUMBER OF LOCAL STATISTICS (L)</div> </div>	GLOBAL (L = 1)	Many GlobalUni	Many / Some GlobalBi	Some GlobalMulti	Some / Limited GlobalMany
	MACRO (L = Macro)	Many / Some MacroUni	Some MacroBi	Some / Limited MacroMulti	Limited MacroMany
	MICRO (L = Micro)	Some MicroUni	Some / Limited MicroBi	Limited MicroMulti	Limited / None MicroMany

Table 6.2: The ability to make comparisons when visualising multiple scale resolutions, scale extents or multiple transformations with increasing numbers of variables (V) and local summaries (L)

from the energy analysts, where known graphs were preferred to abstract and minimalistic representations (Section 3.5). The use of matrices are proposed when V is greater than 2 as this is a compact, space filling visualisation method for multivariate comparison of correlation (as discussed in Section 2.4.2). Colour encoding for global values is proposed when V = Many, yet more detail of the correlation (through other means of visualisation such as scatterplots) can be shown when V = Multi or less. Asymmetrical matrices are suggested for representing both the statistical and spatial views concurrently, e.g. scatterplots and correlation maps¹ in the same matrix, or one side of the matrix showing un-transformed data compared to transformed data. An alternative to asymmetrical matrix is the use of juxtaposition, where two matrices are placed side by side for comparison. This could be useful for the comparison of two SEs, e.g. the same variables presented for London in comparison to Manchester.

An aim in vPSA is to enable the viewer to visualise as much of the complexity as possible (Sedlmair et al., 2014); however, the ability to make visual comparisons at the MicroMany level ranges from limited to none, as expressed in Table 6.2. Therefore, visual representation options are not proposed for this cell but instead shown an ‘arrow up’ or ‘arrow left’ to MacroMany or MicroMulti. These two cells also have alternative options for

¹The use of maps in a matrix is inspired by some experimental visualisation by Aidan Slingsby at the giCentre, City University London

6.2. POTENTIAL VISUAL REPRESENTATIONS OF THE FRAMEWORK

		← FILTER OR AGGREGATE TO REDUCE V →			
		← INCREASE NUMBER OF VARIABLES (V) →			
		V = Uni	V = Bi	V = Multi	V = Many
↑ FILTER OR AGGREGATE TO REDUCE L ↓	GLOBAL (L = 1)	Histogram or Dot/Box Plot <i>Map (Choropleth) of Raw Values</i>	Scatterplot <i>Pair of Maps or Difference Map</i>	Scatterplot Matrix <i>Series of Maps or Difference Maps</i>	Colour Encoding <i>Colour Encoding</i>
	MACRO (L = Macro)	Series of Histograms or Dot/Box Plots <i>Map (Choropleth) of L Values</i>	Scatterplot coloured by L Values <i>Pair of Maps or Correlation Map</i>	Scatterplot Matrix coloured by L Values <i>Correlation Map Matrix</i>	<div> <div>Fewer Data Items: Filter (AE) / Aggregate (AR or Vis)</div> <div>High Data Density: Saliency Threat</div> <div>More Pixels</div> </div>
	MICRO (L = Micro)	Dot/Box Plots or Colour Encoding <i>Map (Choropleth) of L Values</i>	Scatterplot coloured by L Values <i>Pair of Maps or Correlation Map</i>	<div> <div>Fewer Data Items: Filter (AE) / Aggregate (AR or Vis)</div> <div>High Data Density: Saliency Threat</div> <div>More Pixels</div> </div>	Impossible Design Trade-Off

Table 6.3: Statistical (top row) and spatial (italics in bottom row) visual possibilities within the framework as V and L increase

visual representation as the amount of data to display is potentially vast and comparison may be limited (see Table 6.2). The need for a decision when visualising the values from these cells is emphasised in Table 6.3 where the given options are: remove data items (aggregate or filter the data), show all the data (yet increase the threat of saliency) and increase the number of pixels (change technology). Data items can be filtered by reducing the size of the SE. This can refer to geographical size, a particular attribute or domain, or a period of time. Alternatively, data items can be reduced by aggregating the data at the StR level (again by geography, attribute or time) or by aggregating the visual encoding on the fly for visualisation purposes and retaining the information for alternative views; for example, combining all data items in one location and representing only the average on the screen. Finally, data items could be reduced for L by using an alternative type of locality, for instance by using partitioning and reducing the number of output values (see Section 5.1). If data item reduction is not chosen for this stage, then all the data can be shown but there may be overlapping problems which can hinder the ability to read all the data or alternatively more screen space can be sought to show all the data at once on larger or multiple screens. These three options for MacroMany or MicroMulti are demonstrated through the visual prototype explained later in the chapter.

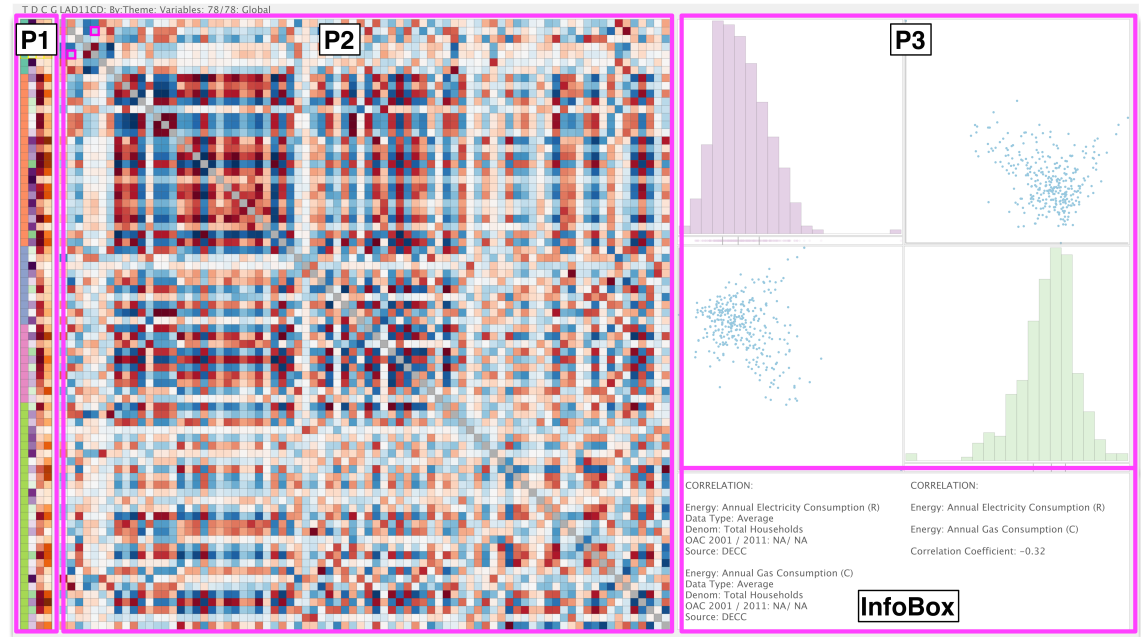


Figure 6.1: The Prototype Layout Panels: P1: Overview, P2: Comparison, P3: Detail. The metadata for the selected variables is shown in InfoBox in the lower third of P3

6.3 Visualising the Framework

The framework can be used as a guideline for designing data visualisations or visual analytic tools for the comparison of multidimensional geographical datasets. In reference to DSR (as explained in Section 2.2.2) the framework can be described as a new artefact (model) containing proposed constructs (types of scale and locality) and methods (the graphic designs and algorithms for the calculations). To realise this new framework in a real-world situation, an instantiation (in the form of an interactive visualisation prototype) of the framework is designed and described in this section within the context of variable selection for energy-based geodemographics.

The literature and previous research in this thesis has shown that the current variable selection process is complex and time-intensive and that there is a need for a more intuitive approach where interactive visualisation can be of benefit. In reference to the parameters discussed for the framework, the SR and SE of the data along with the variable distribution ($V = \text{Uni}$), multivariate correlation ($V = \text{Bi, Multi and Many}$) and geographical variation and sensitivity to scale ($L = \text{Global, Macro and Micro}$) are all shown to be important factors when selecting variables for clustering. Statistical methods used in the process, such as standardisation, transformation and clustering and now the inclusion of locality, are also important for the creator to understand. As there are currently limited representations of SR and geographical variation in the geodemographic literature (see Section 2.3) the instantiation concentrates on these two aspects.



Figure 6.2: LAD maps from the prototype of three variables with differing values of Global Moran's I

Design decisions for the prototype are explained, illustrated and justified in this section. To accompany the explanation of the visual design a video² demonstrates the interactive comparison capabilities. Requirements and ideas for the design were drawn upon from the research undertaken in Chapter 3 and Chapter 4 where creative and novel designs for the energy industry are investigated. In Section 6.3.4 prototype is verified using the context and terminology of the framework prior to validation which continues in the following chapter.

6.3.1 Prototype Design Decisions

This section describes and justifies the prototype design decisions in relation to the research questions and the visualisation literature described in Chapter 2. The default SR for the prototype (as shown in the video and most of the figures in this section) is LAD as this is the most detailed SR with locality included and therefore best represents the full framework.

6.3.1.1 Layout

In the prototype, increased detail of data can be read from a general overview on the left through to fine detail on the far right. The general layout of the prototype was influenced by an infovis design by Al-Awami et al. (2014), which also has a high-level overview, a medium amount of detail at the comparison level and a very detailed view. The three views in the prototype are described as panels: Panel 1 (P1): Overview (left) – re-ordering, identifying and selecting single variables from a list of all variables (Many), Panel 2 (P2): Comparison (middle) – selecting and identifying pairwise variable correlation when comparing Multi to Many variables (adaptable) and Panel 3 (P3): Detail (right) – showing single (distribution) or pairwise (correlation) comparison and allowing for the exploration

²this is available in high quality in the digital appendix submitted with the thesis and online at: <http://vimeo.com/112182748>

of individual data items across views. These three panels are illustrated in Fig. 6.1. The visual representations shown in P2 adapt when filtering is implemented, depending on whether multi or many variables are compared at one time. The visuals in P2 and P3 are adapted based on the decision for statistical or spatial views and whether global or local data is being investigated. These adaptable visuals, transitions and interactions are shown throughout the accompanying video³, particularly in the first 2.5 minutes. In all three panels, comparison is possible and the visual variables (position, order and colour) are all equally important. Certain design rules were implemented in the development and these are described below along with the detail of each of the panels.

6.3.1.2 Panel 1: Overview

Each variable from 1 to 78 is positioned in juxtaposition down the screen with four columns across the screen representing four overview variable values using colour. Above these columns are the letters ‘T, D, C, G’ which refer to the four terms: ‘Theme, Distribution, Correlation and Geography’. These are the four overview dimensions chosen to represent each of the 78 variables at a global level, as they have been shown to be important to variable selection for geodemographics (see Section 2.4).

Firstly, the variables have been grouped into 5 domain based themes: Energy (dark green), Demographic (orange), Household Composition (blue), Housing Type (pink) and Socio-Economic (light green). The justification of all the colour schemes is explained in detail in Section 6.3.1.7. Theme is the default dimension used for ordering, as the energy variables are of most interest to the case-study and this keeps similar variables together. This default order is not only relevant to P1 but is directly reflected in the order of P2.

The other three columns are numerical and are represented by three global statistics: Firstly, distribution is represented by skewness, coloured using a diverging scheme of green (negative) to purple (positive) as shown in Fig. 6.19. The second, variance of correlation was chosen as the global statistic to represent correlation as it reveals how varied the 78 correlations are (the individual variable correlations are also shown in P2), with the darker the red the higher the variance. Geography is then represented in the final column using Moran’s I, for an indicator of how geographically clustered or dispersed the variable is (see Section 2.4.3 for more detail), again this uses another diverging scheme (see Section 6.3.1.7 for more details).

Although Moran’s I has a potential range of $-/+ 1$ the results for the 78 variables range from 0 to 0.38 at the LAD level and -0.15 to 0.57 at the NUTS2 level. This demonstrates that at these scales the variables cannot be described as extremely auto-correlated. As the variable results mainly fall above zero there is a tendency towards the variables being more

³this is available in high quality in the digital appendix submitted with the thesis and online at: <http://vimeo.com/112182748>

clustered than dispersed. Although the range of Moran's I for the 78 variables is limited, the distinction between variables can be seen when using the global value in combination with the distribution maps shown in P2 and P3. In Fig. 6.2 geographical patterns can be identified from differing levels of Moran's I; 'travel to work by public transport' shows only two clusters that of London and the rest of the country, 'single status' reveals minimal clustering around the coastal regions and 'full time student households' has a more random pattern with high values dotted throughout the country – which are very likely to be the university towns of England. These observations indicate that the Moran's I value may be of use in the context of variable selection and therefore it is implemented as part of the prototype design.

As well as re-ordering the variables, the number of V can be decreased and increased using the '-' and '+' keys. In P1 the variables removed from the visual analysis are shaded, while those still in view remain vivid. The removal of V in P1 affects the visual representation of P2, explained in the following section. Variable names and values can be identified through hovering over the cells and selected variables (those shown in P3) are identified in P1 with a thick yellow outline.

In summary, the four columns in P1 represent four generalisations of each variable which are key to the variable selection process and these form four ways of ordering the variables in P1 and P2, which are directly linked by variable order. In terms of visual design, the position of a variable in P1 is important, along with the order and colour of each cell which represents a value.

6.3.1.3 Panel 2: Comparison

The layout of P2 is in the form of a matrix. In the default view shown in Fig. 6.1 this consists of 78 variables by 78 variables, where the order of the columns is identical to the order of the rows. In the default view the colour of the cell represents the global correlation coefficient value for the pair of crossing variables. Correlation is consistently encoded using the blue (negative) to red (positive) continuous diverging colour scheme illustrated in Fig. 6.19 and explained and justified in Section 6.3.1.7. The position of the variables in P2 relates to the ordering of variables in P1 with Fig. 6.3 illustrating the order of the global correlation coefficient matrix for each of the four overview statistics of P1, each time ranging from max to min values (this also be can be seen in the video at 0m:18s-0m:27s). Re-ordering allows the user to identify which variables are in need of investigation, i.e. heavily skewed variables, strongly correlated variables or variables with limited geographical distribution.

Colour, order and comparison are all important in P2, yet it is the use of filtering which links the prototype directly to the framework. In P2 the 78 variables can be filtered

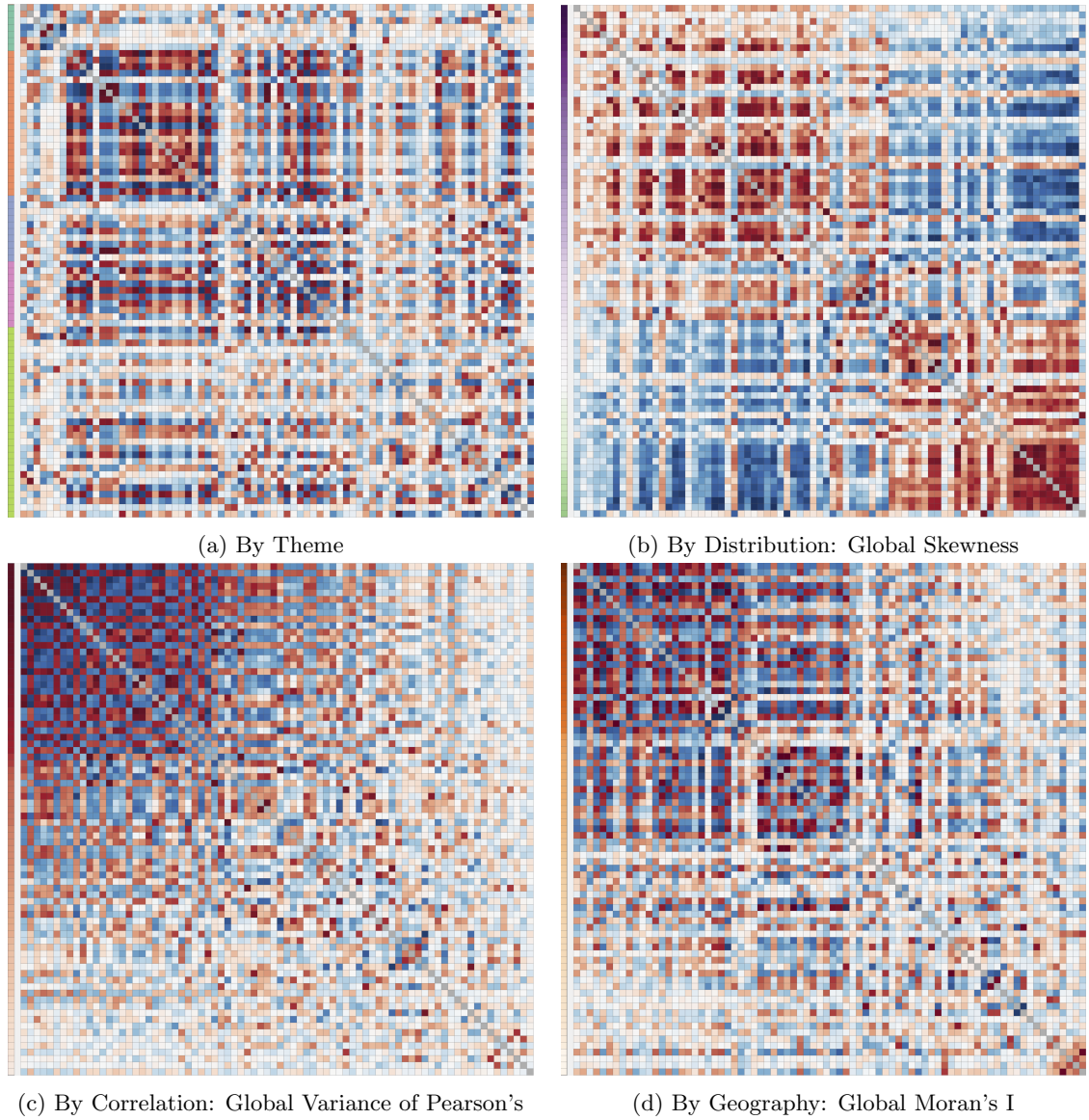


Figure 6.3: Reordering the GlobalMany view in P2 using the four reordering options in P1: Theme, Distribution, Correlation and Geography

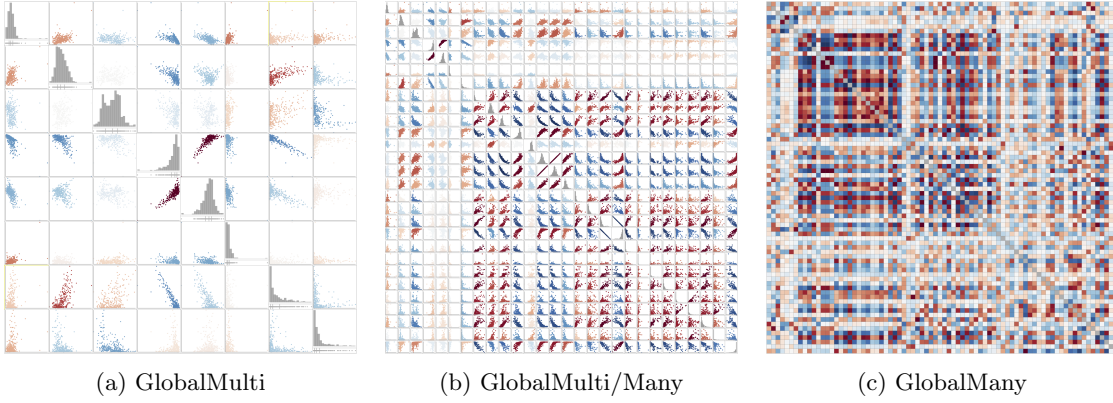


Figure 6.4: The visual representation of P2 showing colour encoded global correlation coefficients values from GlobalMulti to GlobalMany

to any number of variables (from 2-78) to compare and with this filtering process comes the ability to view more detail and change the visual representation. This process of filtering is demonstrated in Fig. 6.4, where the colour used to encode the global correlation coefficient is used to colour the scatterplots to double encode the correlation of the variable by scatterplot shape and direction as well as colour.

As yet, in the explanation of P1 and P2 only the visual representation of the first row of the framework, that of global statistics where $L = 1$, has been reported. To incorporate the next two rows of the framework, the matrix of P2 is utilised to illustrate the spatial and the statistical views of the local (Macro and Micro) correlation values simultaneously, through the use of an asymmetrical matrix. The spatial views are shown below the diagonal and the statistical views are shown above (see Fig. 6.5). Here, the correlation maps and the scatterplots are coloured by the local correlation coefficients. A default N for the chosen SR is selected automatically, with $N = 25$ for LAD and $N = 5$ for NUTS2. This can be changed with a keystroke ('L') to allow the sensitivity in variable geographical variation to be explored depending on the calculation of locality. For the demonstrative data, Fig. 6.4a-6.4c shows how the map representations change when varying the nearest neighbours from 25 to 50 to 100 for the local correlation and skewness. The example in Fig. 6.7 uses the same variable pairs of electricity and gas consumption, as demonstrated in the previous chapter in Fig. 5.2). Re-ordering the matrix when the Macro or Micro data is visible reorders the variables in the same way as globally. Detail of the geographical variation of the variables is evident using re-ordering with locality data, as shown in Fig. 6.8. The video shows the inclusion of local geographical variation in the interactive prototype in detail (1m:28s-3m:10s).

As more variables are added to the comparison view (P2), the use of aggregation enables the spatial and statistical local views to still be represented, albeit at a Macro scale (in this case, P3 still represents the Micro scale), until the screen pace runs out, at

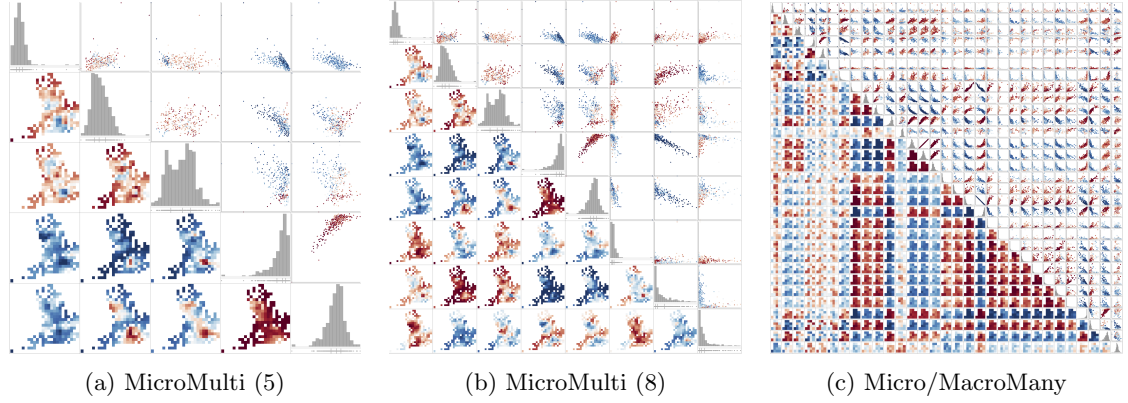


Figure 6.5: Asymmetrical matrix of P2 representing the local spatial and statistical visual representations of $V = \text{Multi}$ with increased spatial aggregation as V increases reducing L from Micro to Macro

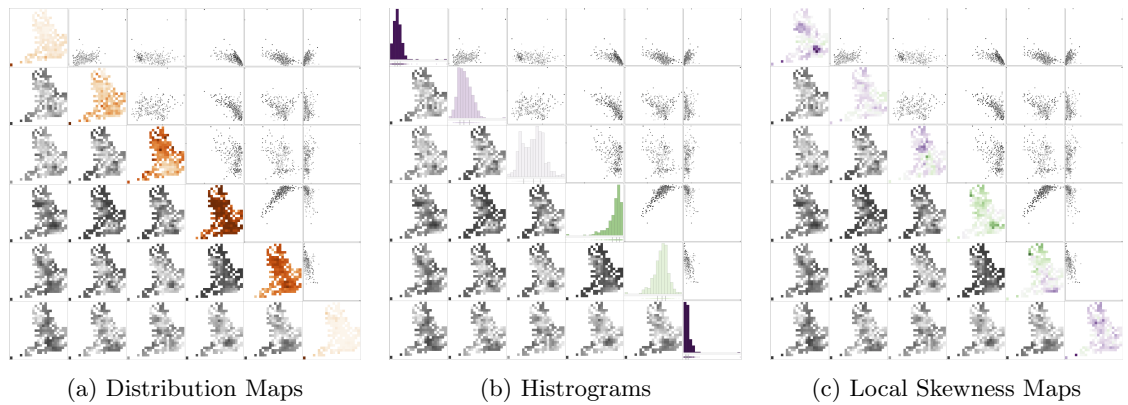


Figure 6.6: Use of diagonal for representing geographical distribution, statistical distribution and local skewness as interchangeable graphics

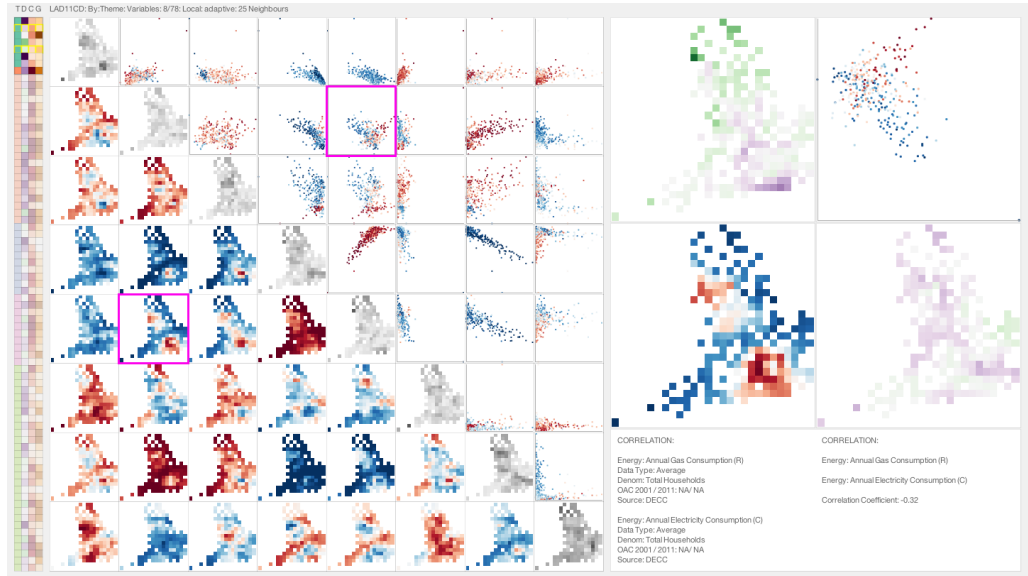
which point the decision has to be made as to whether to aggregate to global or to filter the variables (as shown in the framework). In the prototype design a maximum number of pixels for the size of the maps is used to demonstrate the transition from MacroMany to GlobalMany. 20 pixels was chosen as the maximum cell size for the aggregated maps (see Section 6.3.1.5) as the shape of the map became unrecognisable at smaller sizes. It is likely that this is still too small to distinguish local patterns at the smallest level. Further research is needed to investigate the readability of the asymmetrical matrix as V increases and matrix cell sizes reduce.

The diagonal of the matrix in P2 in both the global and the local views is redundant in the case of correlation so it has been used to represent the variable distribution – a histogram featuring a dot plot underneath, a variable distribution map or a local skewness map are all interchangeable in the central diagonal position in order to investigate the local variation of the distribution (see Fig. 6.6 and video: 1m:47s-1m:58s). When $V = \text{Many}$, the cell is colour encoded to represent the global skewness value (see Fig. 6.20d). Distribution uses a sequential colour scheme in the case of the distribution map and the skewness diverging scheme is used for the cell colour, the histogram and the skewness map (as shown in Fig. 6.19). When $V = \text{Many}$ the skewness value shown in the diagonal duplicates the skewness value shown in the second column of P1. As V is reduced the detail of the distribution is revealed, as this is an important aspect of the variable selection process.

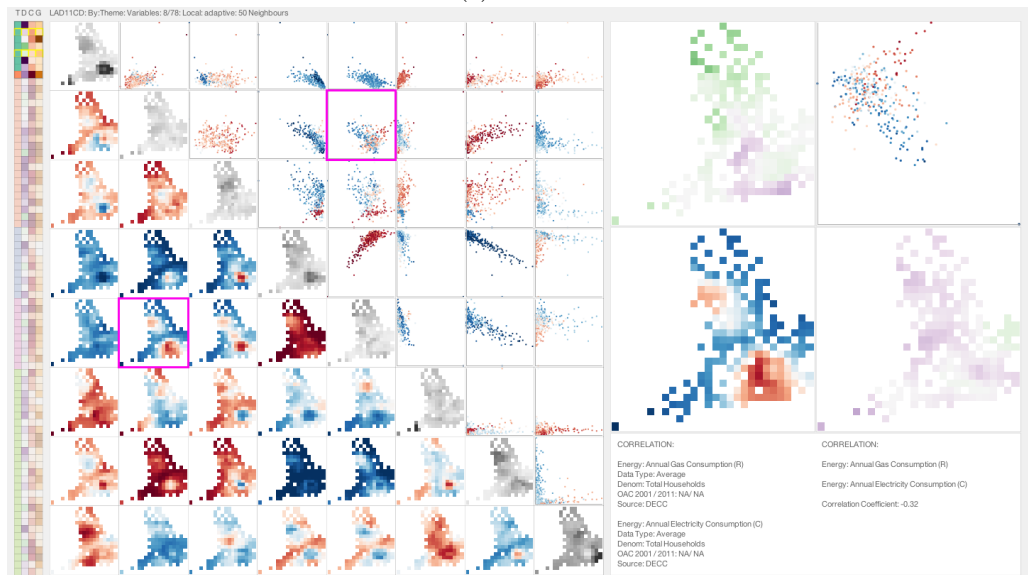
In addition to the detailed correlation views, many of the visual designs shown in Section 6.2 can be displayed in P2. For example SR is demonstrated using superposition containers named ‘scale mosaics’. This technique is discussed in Section 6.3.1.6. As with P1, the important visual design features of P2 are position (both juxtaposition and superposition) as variable relationships are shown, colour which represents values and order which aids the comparison process.

6.3.1.4 Panel 3: Detail

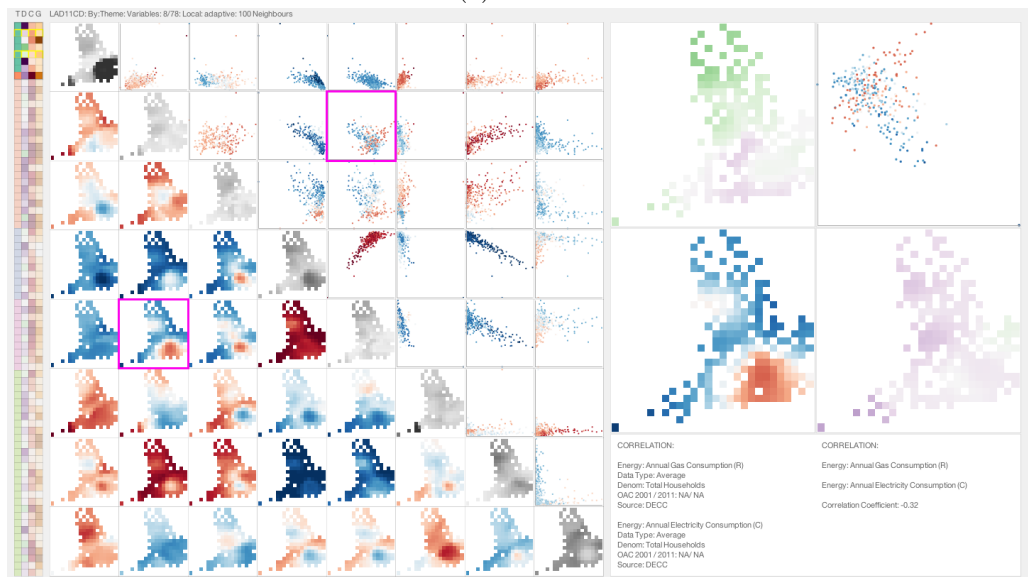
Unlike P1 and P2, P3 shows the detail of one or two variables rather than an overview of multi or many variables and therefore represents the first two columns of the Framework (shown in Tables 6.1, 6.2 and 6.3) of $V = \text{Uni}$ and $V = \text{Bi}$. The design rules of P3 reflect the visual possibilities expressed in Table 6.3. When $V = \text{Uni}$, activated by clicking on the diagonal cell of P2, the variable distribution is shown in detail in P3 as a large histogram (coloured by skewness) and a distribution map to display the geographical distribution of the variable, with an option to change this distribution map to a local skewness map (when Macro or Micro values are shown). An additional option for this view is to show the effect of the log transformation on the skewness of the distribution. Fig. 6.9 shows that



(a) $N = 25$



(b) $N = 50$



(c) $N = 100$

Figure 6.7: Demonstrating the sensitivity of the locality calculation as N in the adaptive moving window approach increases from 25 to 50 to 100 neighbours.

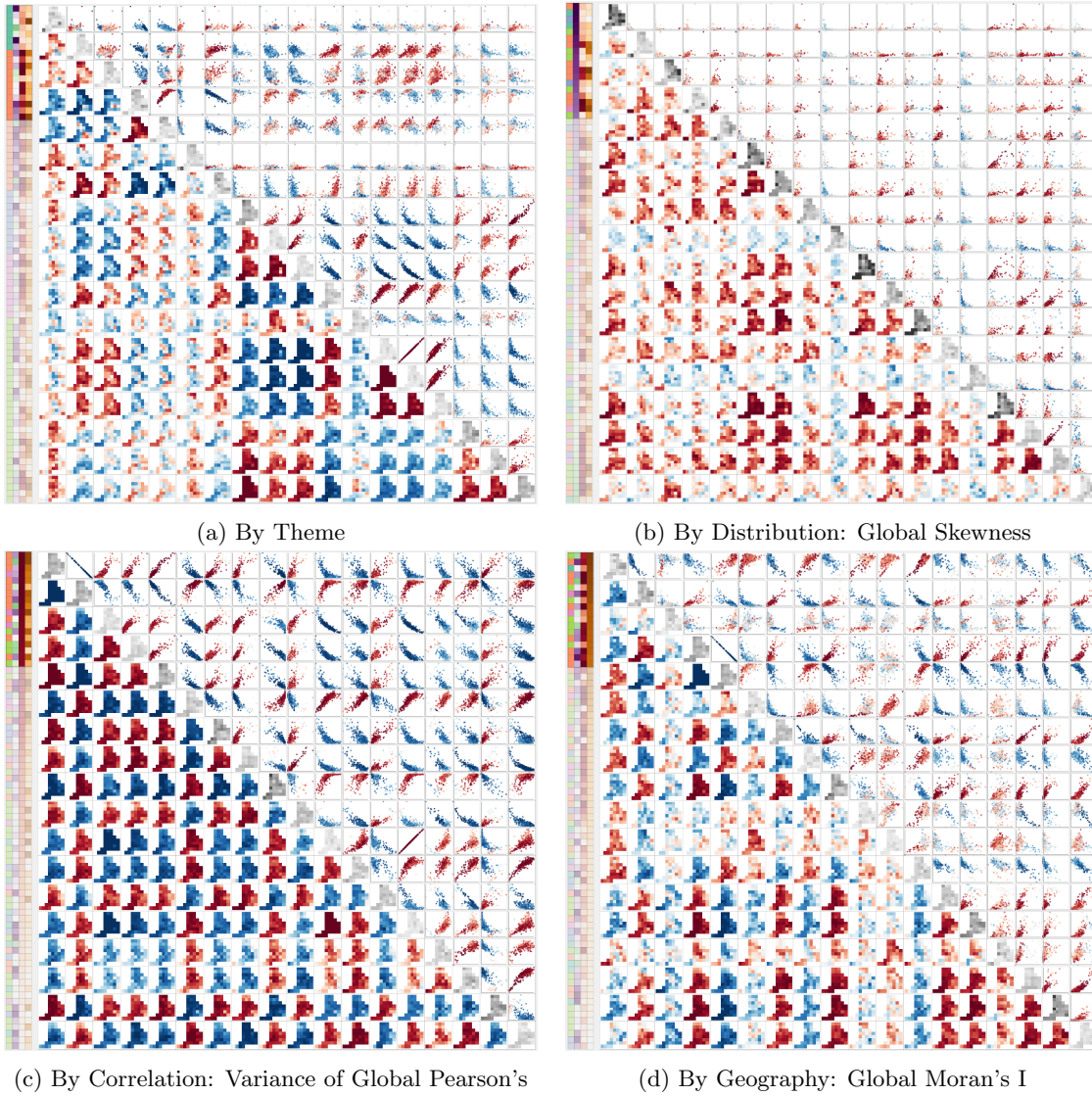


Figure 6.8: Reordering the MultiMacro view in P2 with the inclusion of local statistics and the asymmetrical matrix using the four reordering options in P1

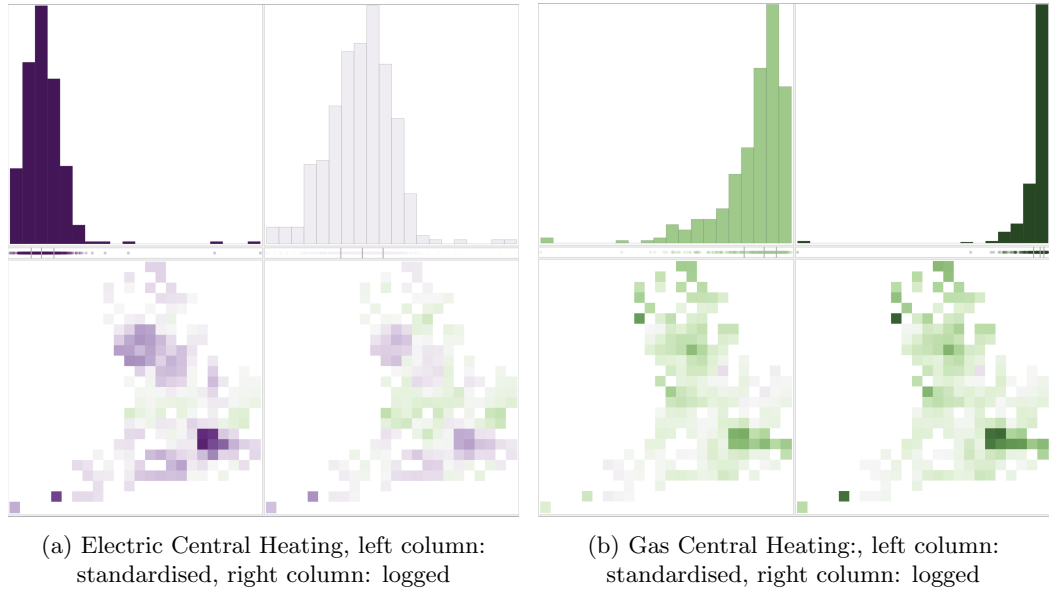


Figure 6.9: Panel 3 showing electric and gas central heating variables transformed via the logarithmic scale, with the original and logged distributions of the variables as histograms (top) and local (adaptive 25 neighbours) skewness maps (bottom)

the log transformation improves (moves towards normal) the distribution for the positively skewed variable but increases the skew of the distribution which is negatively skewed. The histograms and their associated dot plots can be shown in P3 and to aid the readability of the scatterplot they can be turned towards the relevant axis, as shown in Fig. 6.10.

For $V = B_i$ (clicking elsewhere in the matrix of P2), pair-wise correlation is investigated. By default two distribution maps are shown in P3, one for each variable of the comparison, together with a large scatterplot, where the data is coloured according to the global correlation value, reflecting the P2 view, as shown in Fig. 6.11. There is an option to change the two maps to two histograms to allow the statistical view to be shown (as shown in Fig. 6.10). When Macro or Micro local detail is shown in P3, the local correlation map is also shown with the two distribution maps and the scatterplot, where the data is now coloured according to the local correlation values. The distribution maps can also be replaced with local skewness maps in this view, as demonstrated in Fig. 6.12.

In P3, position is kept constant depending on the visual representation shown (i.e. scatterplots and maps follow the diagonal, as represented in P2). Colour is important and is kept constant throughout the three panels. P3 is used to inspect the detail of the variable and the comparison of two variables. While this is notably useful for the inspection of the variables in the global view, the inclusion of locality shows the detail of the geographical variation of the variables and interesting patterns can be identified through the combined views of P2 and P3, for example the distinct patterns of the two variable pairs shown in Fig. 6.13 and Fig. 6.14. Fig. 6.13 shows that a mid-level positive global correlation consists of both highly positive and highly negative correlations. The

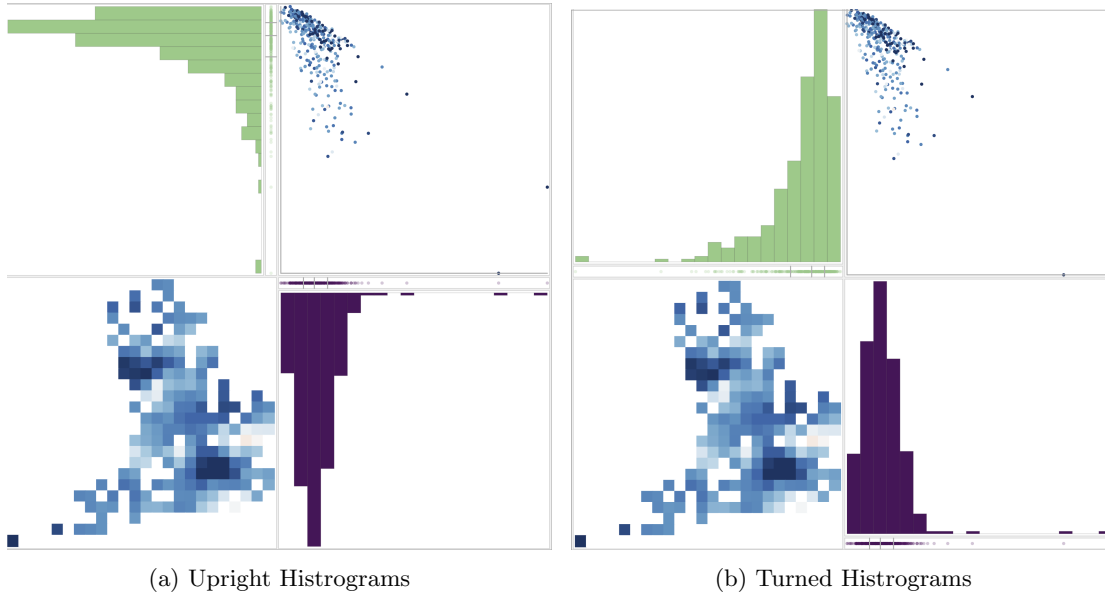


Figure 6.10: Panel 3 showing upright and turned histograms: ‘gas central heating’ (top left) and ‘electric central heating’ (bottom right). Local correlation coefficient (adaptive 25 neighbours) is shown in the map and the scatterplot

map view shows that the most of the negative correlation only appears in the area around London. Fig. 6.14 demonstrates that a variable with no global correlation can consist of extreme positive and negative values and these may be geographically variant. These two examples are shown interactively in the video (2m:45s-3m:05s).

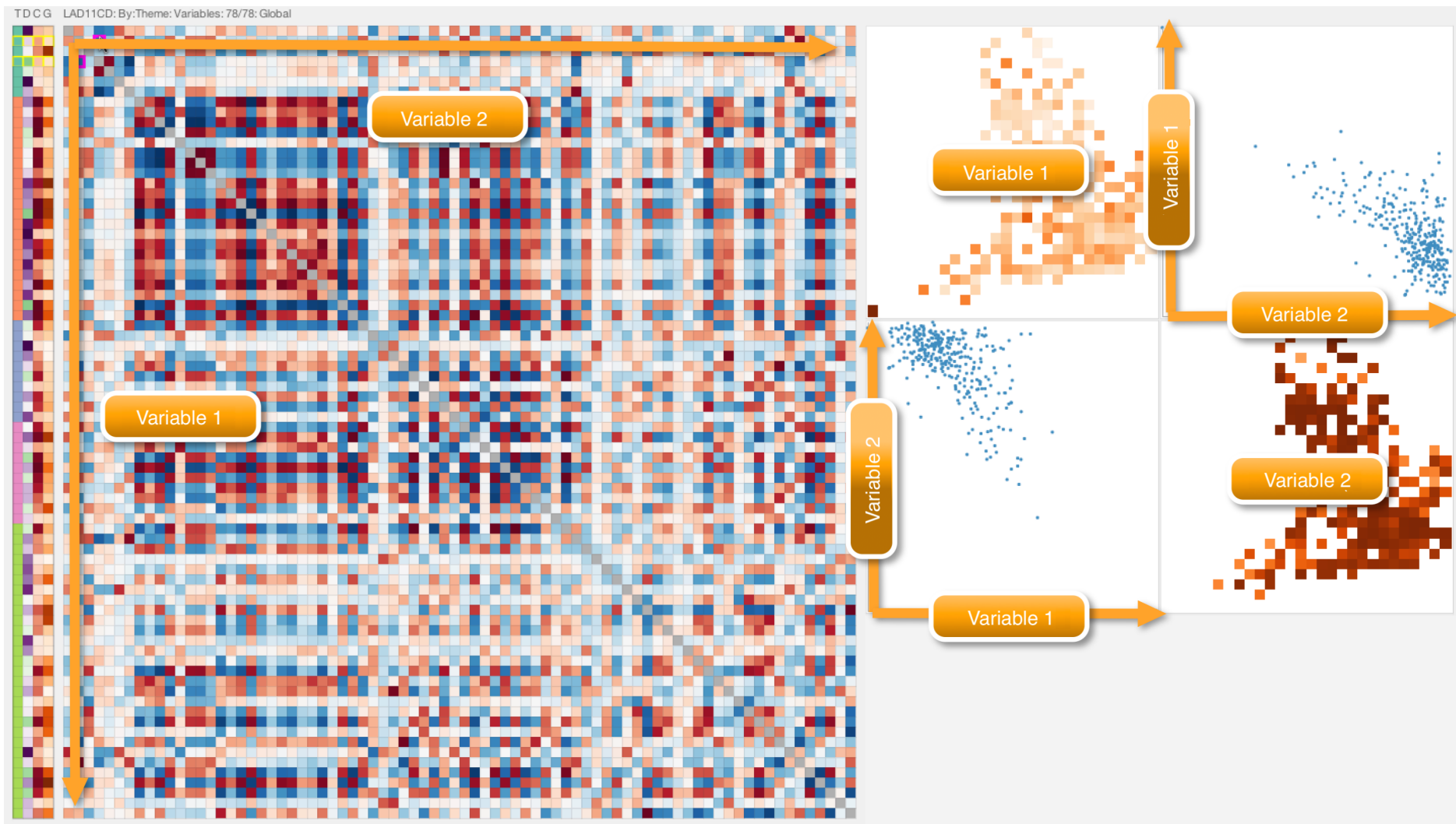


Figure 6.11: Default view for global values with two distribution maps shown in P3, one for each variable of the comparison, together with a large scatterplot coloured by the global correlation reflecting the P2 view

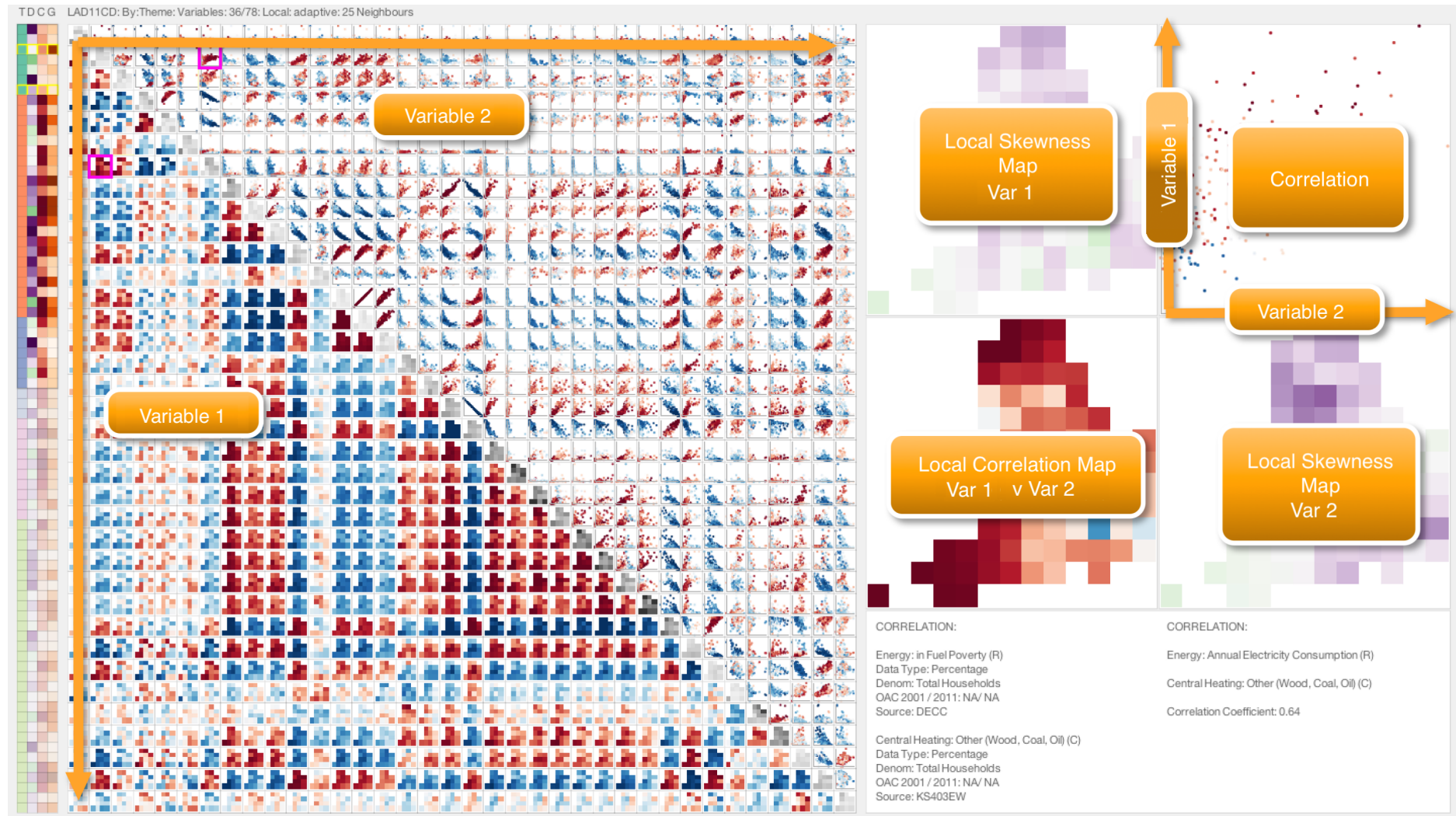


Figure 6.12: When local detail is displayed the local correlation map is shown in P3 with the option to display the distribution map, histogram or local skewness map (shown here) for both variables along with the scatterplot

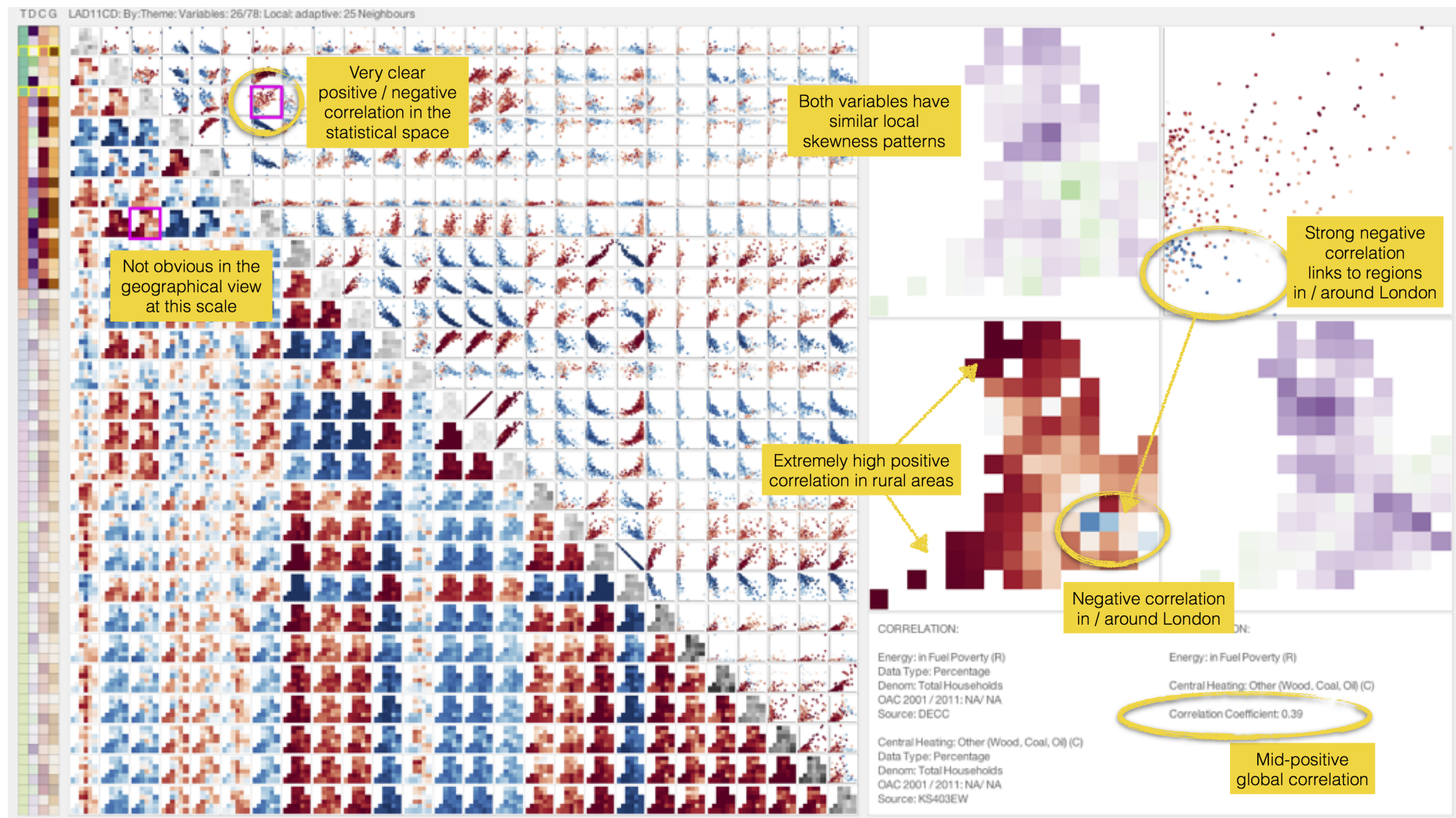


Figure 6.13: Variable pair which shows a distinct difference in the local statistical values for London compared to elsewhere in the country. An example of the patterns which are identifiable when locality is included in variable selection and the geographical and statistical views are shown concurrently (in P2 and P3)

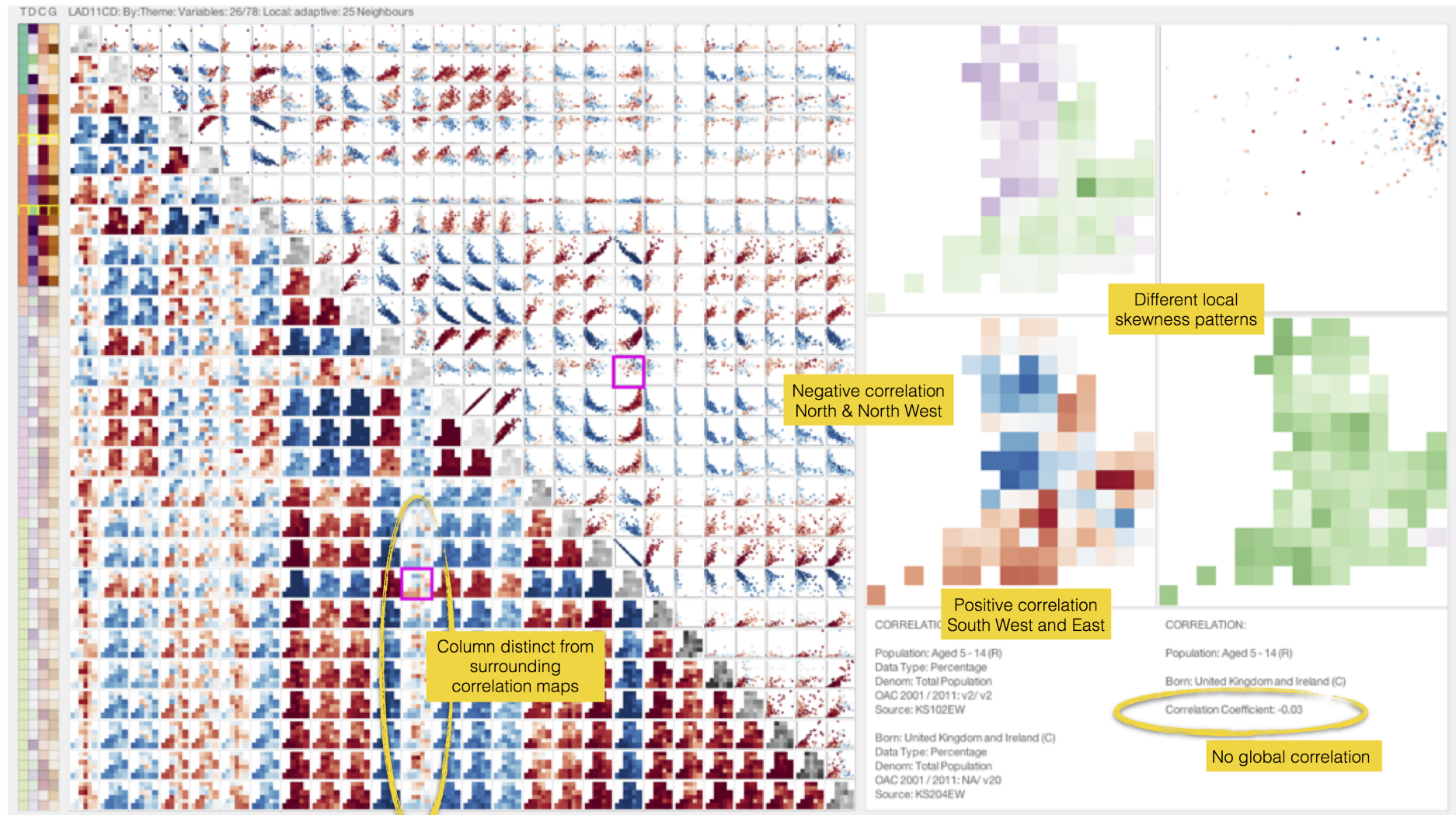


Figure 6.14: Variable pair with no global correlation yet clear statistical and geographical differences when locality is included. An example of the patterns which are identifiable when locality is included in variable selection and the geographical and statistical views are shown concurrently (in P2 and P3)

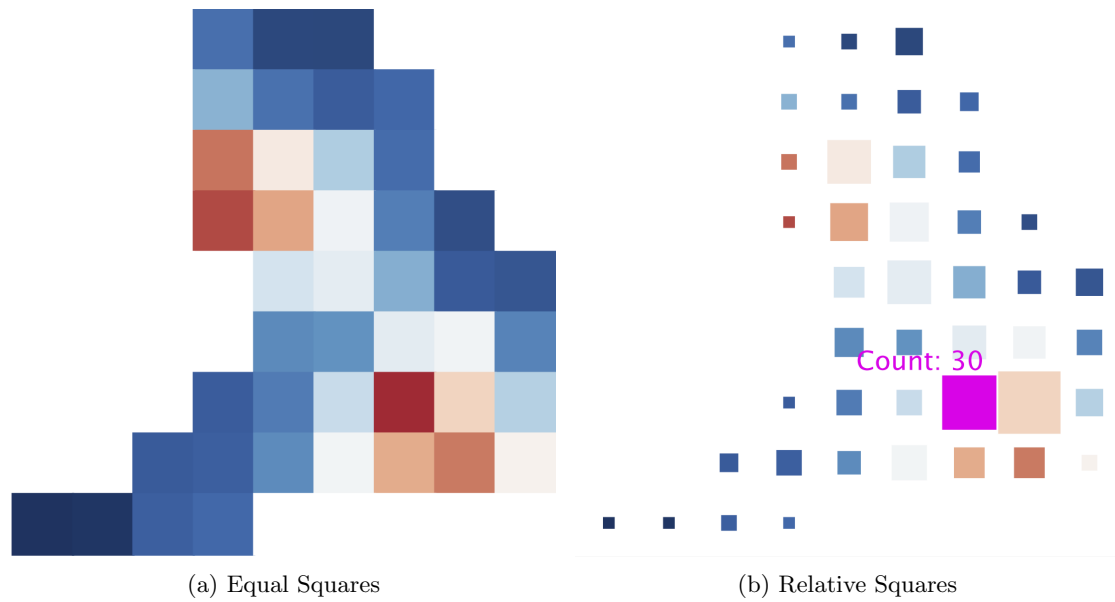


Figure 6.15: Alternative map views with squares of equal size or relative to the number of regions located within

6.3.1.5 Spatial Aggregation v Data Density

In order to represent the geographical variation of the data in a map view which can be visually compared over multiple variables, the raw data items from the output scale are aggregated using a grid square approach to create the square map view. These grid-like raster maps (for distribution, correlation and skewness) are not 100% representative of the data as the same size squares are used throughout and this does not represent the underlying data as there are more regions in the densely populated areas than the rural areas. As many maps are shown at one time in juxtaposition it is helpful to retain this equal square view. An additional map view can be shown in the prototype (by pressing ‘M’) in which the squares represent the number of regions within them (as shown in Fig. 6.15 and video: 1m:57s-2m:08s). This represents the underlying data and can be used to more accurately investigate patterns based on density of the population.

While an understanding of the geographical extent of the area is useful for interpreting the data, it is a method of representing geographic variation rather than a means of identifying individual places. The degree to which geographic variations occur and can be visually detected depends upon the resolution. As V increases, the available screen space is reduced for each map and therefore the more difficult it is to identify local geographical variation.

In terms of data reduction for visualisation, the map view represents aggregated data, while the statistical view does not. The scatterplot shows all the output scale data items; even the thousands of regions for LSOAs can be shown (see Fig. 6.16), although it becomes difficult to discern them. Data items are increased or decreased by changing the SR, the

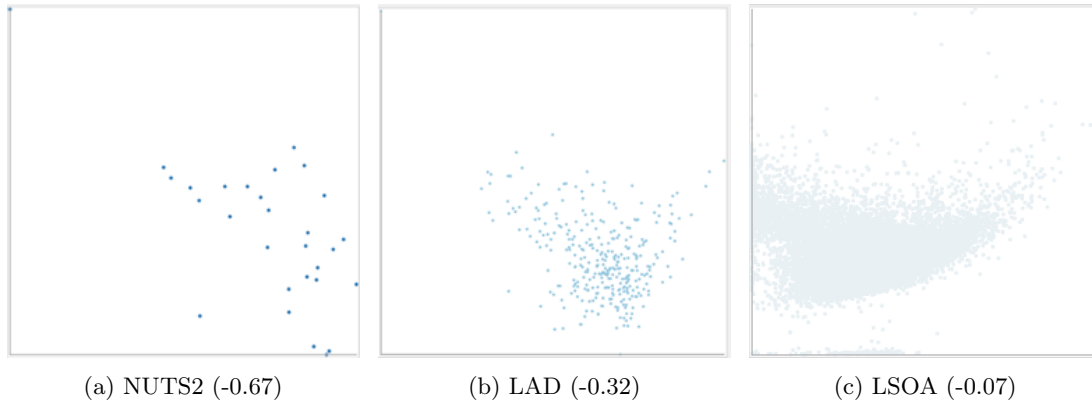


Figure 6.16: Increased number of data items per scatterplot when alternative SR (NUTS2, LAD and LSOA) are shown. Comparing ‘gas’ (x-axis) and ‘electricity’ (y-axis) consumption. Data items are coloured by global correlation coefficient shown in brackets

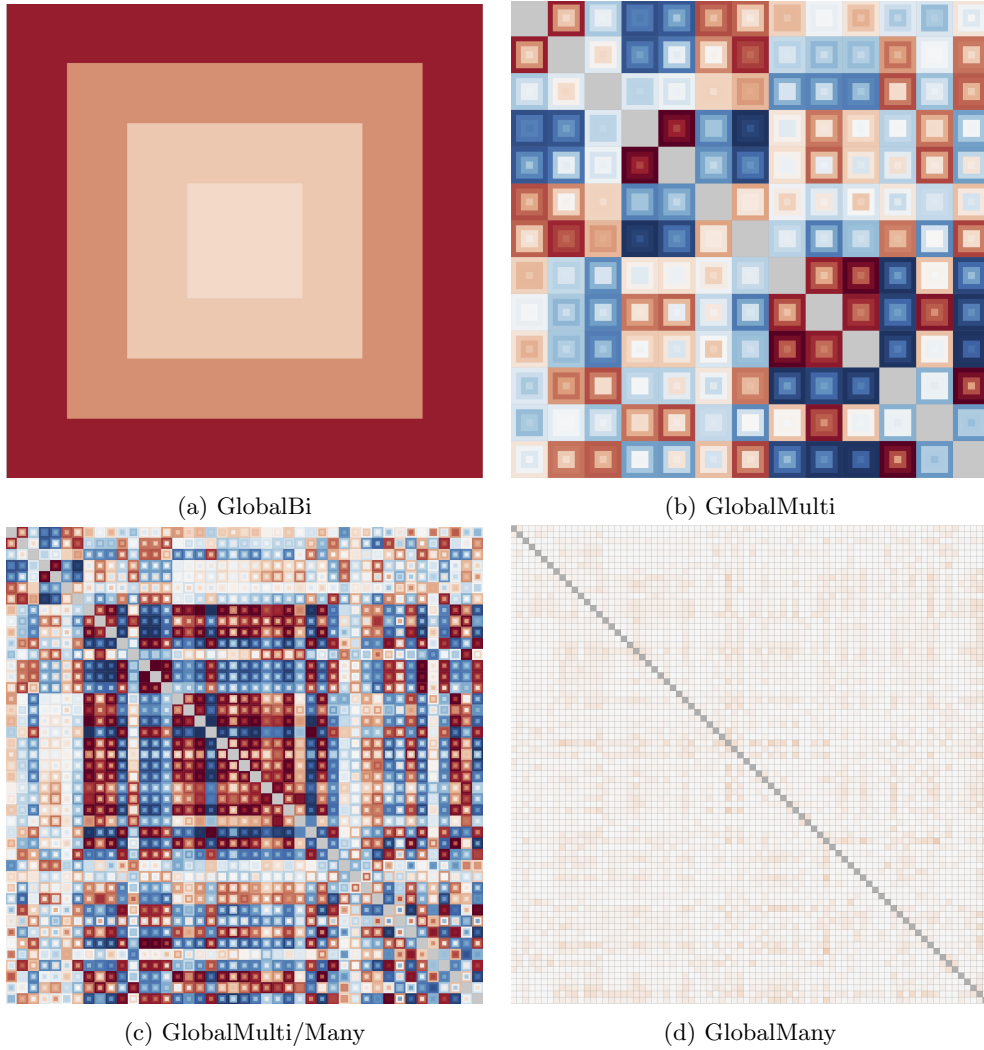


Figure 6.17: Scale mosaic view with superimposed colour encoded global correlation coefficients (CC) for four SR from $V = \text{Bi}$ to $V = \text{Multi}$. At $V = \text{Many}$ the variance of the four CC values is encoded, the darker the more varied

number of variables or the inclusion of locality. As data item quantity increases so does the data density on the screen and with this a threat of saliency. The more items, the more difficult it is to distinguish the scatterplot points. Whilst aggregation can be used for scatterplots (as discussed in Section 2.4.2), instead opacity is used in the scatterplot view to determine over-plotting and the radius of the points can also be reduced by the user. As more data items are added (as scale is increased) the detail becomes difficult to interpret; however, it is the correlation which is critical to the context, rather than the location of individual data items and this can be seen from the shape of the scatterplot. The use of colour to represent the global correlation coefficient also aids the interpretation.

The difficulty in representing all data points is particularly evident when including locality and this is expressed in the framework in MacroMany and MicroMulti (see Fig. 6.3), where user decisions are needed as to whether to overplot, filter, aggregate or to change technology to increase the number of pixels available. Two different decisions were made for the prototype; a decision to aggregate the geographical view, but increase the data density of the statistical view (as demonstrated in Fig. 6.5 and shown in the video: 2m:11s-2m:30s). This decision is based on the fact that similar values are likely to be close to each other on the map (according to geodemographic research see Chapter 2) and hovering over an area of the map in P3 (see interaction in Section 6.3.1.8) can aid the understanding of introducing geographical variation into multivariate comparison by identifying how these data items cluster together in the statistical space (in the scatterplot).

6.3.1.6 Scale Mosaics

The scale mosaic shows the global correlation coefficients or skewness of the chosen four SR are shown as differing sized squares within one matrix cell or mosaic tile. By default, the outer square representing the largest geographical regions, in this case NUTS2, and the central square representing the smallest geographical regions, in this case OA. This is reversible to show the inverse pattern. The scale mosaic view is shown in Fig. 6.17, where the visual representation adapts as variables increase from a pair through to many, where the global value for the variance of correlation is used in the final $V = \text{Many}$ representation. It is possible to investigate the sensitivity of scale through the use of the scale mosaic view, for example the variable pair shown in Fig. 6.18 for each of the four SR in the dataset (and interactively in the video from 3m:15s-4m:15s). This view and the patterns identified are explored in more detail in Section 7.3.2. Additional designs for mosaics views for different types of scale are also discussed for future work in Section 8.1.

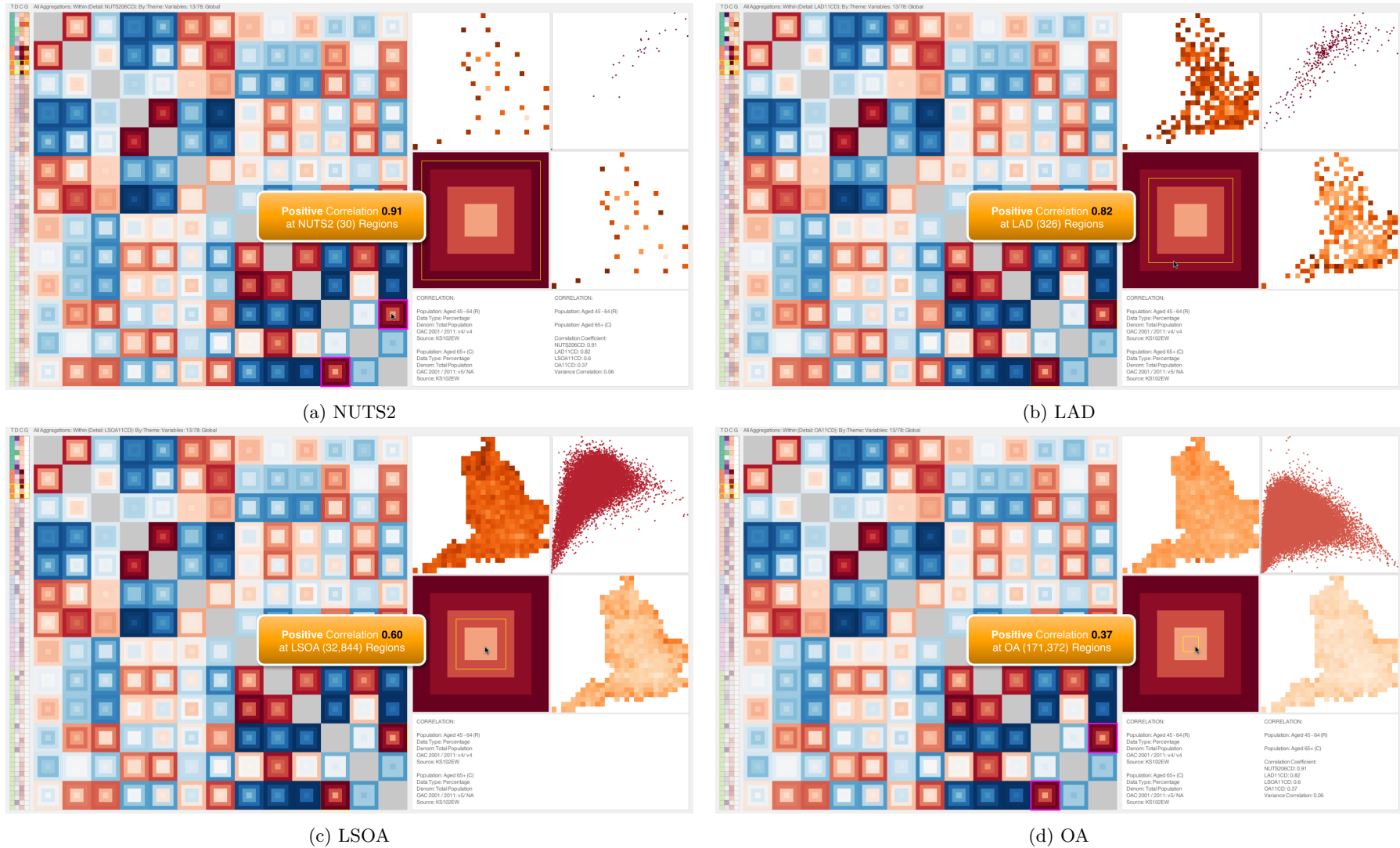


Figure 6.18: Scale mosaic view enables the sensitivity of SR to be identified for the global correlation and skewness values for variable pairs. The prototype allows the values to be investigated for each of the four SR: NUTS2, LAD, LSOA and OA

6.3.1.7 Colour

The consistent use of particular colour schemes in the prototype has been mentioned during the explanation of the three panels. A number of different colour schemes (see Fig. 6.19) have been chosen for particular reasons and these are explained as follows.

The visualisation of data in the form of a map has a number of specific rules related to the type of data values being shown. Sequential, diverging and quantitative data must all be represented using the right type of colour scheme. ColorBrewer (Harrower and Brewer, 2003) schemes were chosen for the prototype, all chosen to be colourblind safe and shown to be good for various display types, in particular monitors. No intervals were chosen for the sequential and diverging schemes because for skewness and correlation it is useful to visualise the subtle changes in values rather than force interval gaps on the range.

In terms of maximum and minimum, the standardised variable values have a range of 0-1 and the correlation coefficient has a scale of -1 to 1 and therefore both can be fully represented through continuous colour range in the prototype (see Fig. 6.19). Skewness, however, can be any number and it is only the sign which distinguishes whether the distribution is negatively or positively skewed. As explained in Section 2.4.1, skewness greater/less than ± 2 can be classed as extreme. For the default SR of LAD, the majority of variables (57 of the 78) are positively skewed, with 28 of the 78 over the ‘extreme’ threshold of 2. Therefore, in order to show a continuous gradient for these heavily skewed variables a threshold of ± 5 is implemented in the prototype (see Fig. 6.19). This threshold could easily be removed and the max and min values represented (or a threshold of 2 could be imposed).

Continued use of the same scheme throughout the three panels was essential for the coherence of the variable being displayed; however, as correlation and skewness both use diverging schemes a decision had to be made in P2 in order to avoid confusion. Not only is the diagonal difficult to see but there is a conflict with the green and red in the two different schemes shown together as this is a problem for green/red colour blindness. Therefore interaction is used and a diverging grey to white scale masks one coloured scheme depending on whether the diagonal is hovered over or not. Detail of this feature is shown in Fig. 6.20, where not only the cells but also the scatterplots, histograms and maps use the grey-white scheme. In the final image (Fig. 6.20d), the use of position of lines as well as grey scale is used to represent the shape of the scatterplot when there are too many variables to show the full scatterplot. Line glyphs are used instead of colouring the whole square on the grey scale as the value sign is not possible to determine in the global view (Fig. 6.20d and video: 0m:56s-1m:04s). Global skewness is duplicated in P1 and therefore the sign for skewness can be quickly determined, whereas the sign can be inferred when visualising histograms and scatterplots for $V = \text{Multi}$.

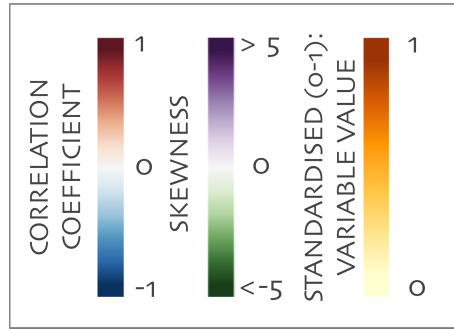


Figure 6.19: Legends of Correlation Coefficient, Skewness and Standardised Variable Value

In addition to these three schemes, a few additional values are represented in the prototype in P1 and each has a colour scheme, yet these are not as dominant or as important to the prototype as the three schemes shown in Fig. 6.19. Firstly the themes have a quantitative scheme chosen from ColorBrewer, using pastel shades so as not to overpower the other colour schemes. The variance of the correlation is represented using the same red colour scheme as in positive correlation (as variance is always positive). The global maximum and minimum are used for the legend threshold in order to show the full scale of the variance. These thresholds depends on the SR shown. Finally, Moran's I is represented by another diverging scheme. Positive values are displayed in orange and negative in purple. Again, the global maximum sets the threshold, which is dependent on scale. As there are very few negative values for Moran's I calculated for this prototype the conflict of both the skewness and Moran's I columns using purple is minimised.

6.3.1.8 Interaction

Interaction with the prototype is implemented through brushing (Monmonier, 1989) and clicking with the mouse and some options are turned on and off through keystrokes. In P1 and P2, hovering over the variable reports the detail of the value that the colour represents, as well as the metadata about the variable (e.g. name, domain, source table, OAC reference) in the InfoBox area of the screen (see Fig.6.1). Ordering of P1 uses the 'O' (order) key, which circulates through all four in sequence. Additional ordering options are also available, including other derived variables not shown in P1.

Clicking on a diagonal cell in P2 triggers the distribution view of the variable to be shown in P3 with the metadata about the variable displayed in the InfoBox. Clicking on any other cell in the matrix triggers the visual representation of the correlation of two variables to be shown in P3. To indicate which cells in P2 and P1 are currently clicked on or brushed, saturated colours (in this case yellow and magenta), which are clearly visually differentiable from the colour schemes, are used. In P3 further interaction can be made by brushing the data items in the scatterplot or map, which reveals the data values and shows the item in the other linked views.

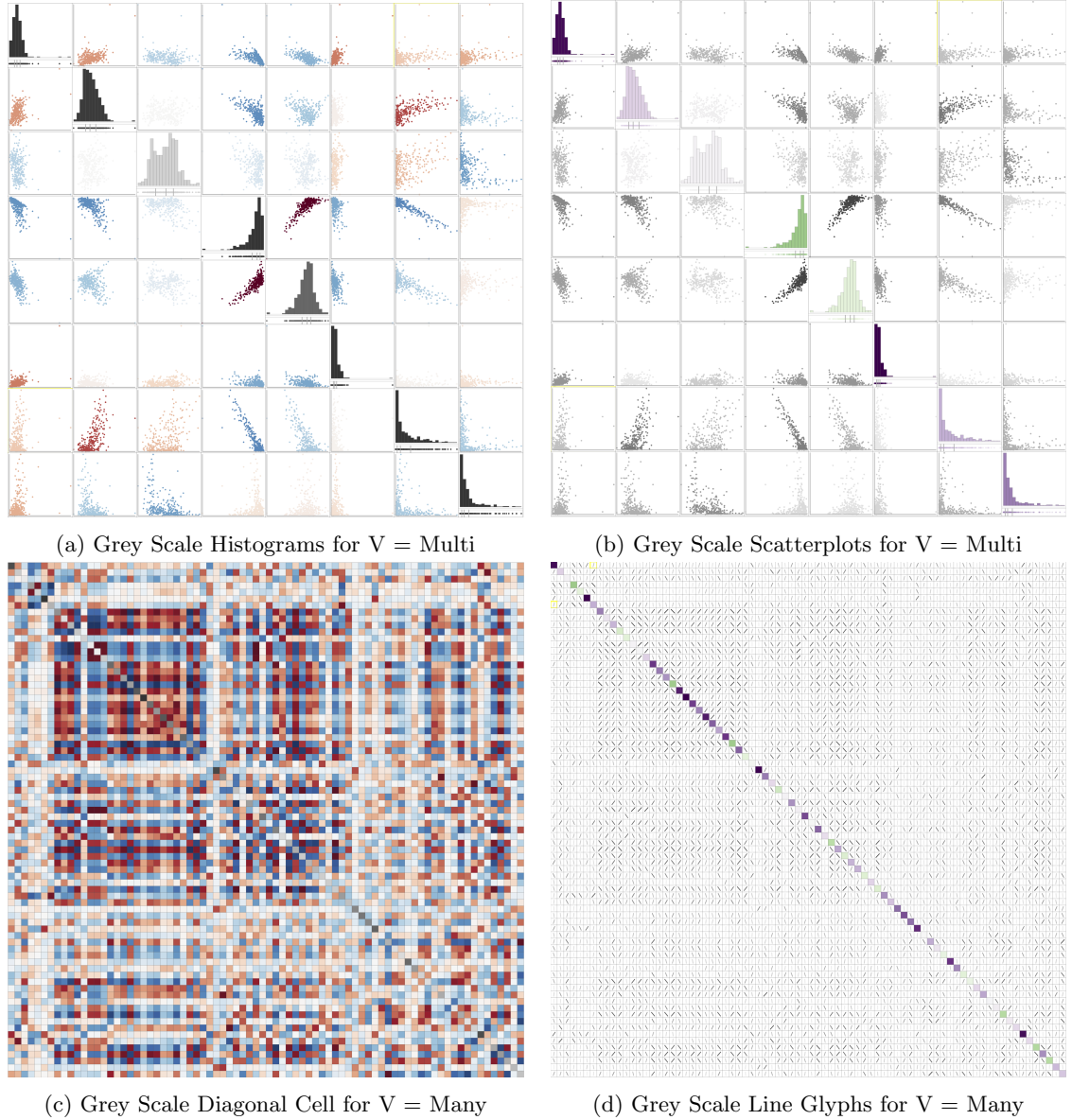


Figure 6.20: Grey-white-grey diverging scheme is used in P2 to avoid confusion of two diverging schemes in the histogram and central diagonal cell (sign inferred from P1) when correlation is being investigated and as the scatterplot or line glyphs when distribution is investigated

Additional options mentioned throughout this section are implemented through different keystrokes for ease of development. Continued research and development is needed in order to implement a fully user friendly application and there are many relevant extensions relating to interaction and usability explained in Section 7.7.

6.3.2 Additional Functionality

Along with the key design decisions mentioned above, various functionality options (using different keystrokes) were added to the prototype to improve usability, these include:

- **Global as Background:** to see how local differs from global it is possible to colour the background of the local statistical views with the global correlation coefficient.
- **Opacity of Data Points:** to reduce the salience threat when data density increases in the scatterplots it is possible to increase/decrease the opaqueness fill colour of the points⁴
- **Size of Points:** to increase the visibility of the individual data items in the scatterplots it is possible to increase/decrease the size of the radius of the points⁴
- **Number of Bins:** to change the representation of the distribution it is possible to increase/decrease the number of bins in the histograms⁴
- **Additional Reordering Options:** in addition to the four options in P1 it is possible to reorder the matrix by kurtosis, median and maximum correlation.
- **Revert to Default Ordering:** after reordering/removing variables it is possible to quickly revert to default ordering and/or add back all variables
- **Show Extreme Values:** to enable problematic variables to be quickly identified it is possible to mask the non-extreme correlation and skewness coloured cells from the matrix (in P2 only). The definition of ‘extreme’ values is taken from the literature (see Section 2.4), with ± 2 used for defining extreme skewness and ± 0.65 used for defining strong correlation. As this option removes the ability to visualise all the statistics for the variable it may also remove values on the edge of the thresholds and could subsequently influence the variable decision or misinform the user, therefore it is ‘switched off’ by default.
- **Data Item Interaction:** the views in P3 are interactive and linked so data items can be identified in the scatterplot and can be seen in the map (this feature is partially implemented and improvements are needed for fully functioning linked views)

⁴This feature is influenced by Mondrian software (Theus and Urbanek, 2008)

- **Filter by Resampling:** to speed up the loading/interaction process it is possible to randomly sample the data shown in the map/scatterplots and increase/decrease the number in the sample (influenced by progressive refinement (Turkay, 2013; Stolper et al., 2014) for visual analytics)
- **Reverse SR in Scale Mosaic View:** reverse the order in the scale mosaic view to make the central square the largest SR instead of the smallest
- **Save Screen:** an image of the current screen (analytical view) can be saved with the essential information as part of the file name to enable it to be reproduced

These additional functions aid the exploration of the variables, investigation of parameter sensitivity and the validation of the utility of the prototype and framework, as discussed in Chapter 7.

6.3.3 Development Techniques

A number of the methods used in the smart home project (Chapter 3) were used to build the prototype, as it needed to be developed quickly and efficiently, meet the key requirement of representing the breadth and essential characteristics of the framework and where possible meet the requirements for variable selection identified during Section 4.2.2. The MoSCoW technique was again used to prioritise requirements and re-prioritisation and re-evaluation took place on a near daily basis, relating to time and development constraints; for example adapting the development plan due to increased run time of the local datasets. All ‘must-have’ functionality was implemented, some ‘should-have’ functionality was included as well as a few ‘could-haves’ which were deemed beneficial and easy to develop. Most of the interaction was not classed as must-have, although some was deemed important for usability of the prototype. Many should-haves and all could-haves were left for extensions (as explained in Section 7.7), as they were not seen as essential for the proof-of-concept. These aspects were linked to user-functionality as well as more complex and advanced features, such as extending the visualisation of scale to include local statistics as well as global.

6.3.4 Referencing the Framework

The use of the dashboard-like design with the three panels enables much of the framework to be visualised at one time, with P3 containing the visuals for $V = \text{Uni}$ and $V = \text{Bi}$ and P2 (and P1) representing $V = \text{Multi}$ and $V = \text{Many}$. A drawback of representing the framework on one screen is that the amount of information that can be viewed in each panel is reduced, in particular the amount of data able to be shown in P2 and therefore the point at which the transition between $V = \text{Multi}$ and $V = \text{Many}$ is implemented. An option to change the view and use more pixels may become relevant when more variables

need to be viewed at one time. Despite the limitations the prototype does represent much of the framework and allows multivariate data to be compared across scale with the inclusion of local geography.

Table 6.4 shows aspects of the prototype within the framework by identifying each visual representation and the panel in which it appears (P1, P2 and P3). Transitioning through the framework is important for the fluidity of analysis and many of these transitions have been implemented. The use of the interaction between the panels allows for the quick shift of visual representation between $V = \text{Multi}$ to $V = \text{Many}$ in P2 as well as $V = \text{Uni}$ and $V = \text{Bi}$ in P3. The flexible threshold between $V = \text{Multi}$ and $V = \text{Many}$ is particularly important to emphasise as this depends on the number of V , the number of L , the number of data items and the size of the screen. These critical transitions of GlobalMulti to GlobalMany and MacroMulti to MacroMany (shown in Table 6.4 as blue arrows 1 and 2) are implemented in P2 and shown in the video at 1m:06s-1m:25s and 2m:11s-2m:30s respectively. When locality is included (i.e. MacroMulti to MacroMany) spatial aggregation and data density (as explained in Section 6.3.1.5) are used to visualise the data. When this aggregation option is depleted (i.e. when V becomes too ‘many’ for the pixels in P2) then the view automatically switches to the global colour encoding, i.e. from MacroMany to UniMany (blue arrow 3).

MicroBi to MacroBi and MicroMulti to MacroMulti (blue arrows 4 and 5) are represented in the prototype in two ways. Firstly, through the changing of data scale, where switching the SR changes the amount of data items being represented. These transitions are shown through the number of data items making up the local correlation maps and in the scatterplots. The SR is altered by using a keystroke. As there are only two variables, MicroBi to MacroBi is represented in P3, while MicroMulti to MacroMulti is shown in P2, where many local maps and scatterplots are visualised. These two transitions in this context are limited by the visual design and the data available. Extensions (discussed in Section 7.7) are possible to improve these transitions and reveal the sensitivity of scale together with the local statistics.

Alternatively these two transitions are represented through the use of spatial aggregation to create the grid-like maps. MicroBi to MacroBi is shown in P3 when switching between very small map squares and very large ones (see Fig. 6.21 and video: 1m:58s-2m:03s). By default the map square sizes in P2 remain static while those in P3 are increased or decreased, as changing the map size in P2 conflicts with the spatial aggregation which occurs when V increases or decreases. There is an option to turn this off, allowing the maps in P2 to also be adaptable and therefore representing MicroMulti to MacroMulti from the aggregation perspective.

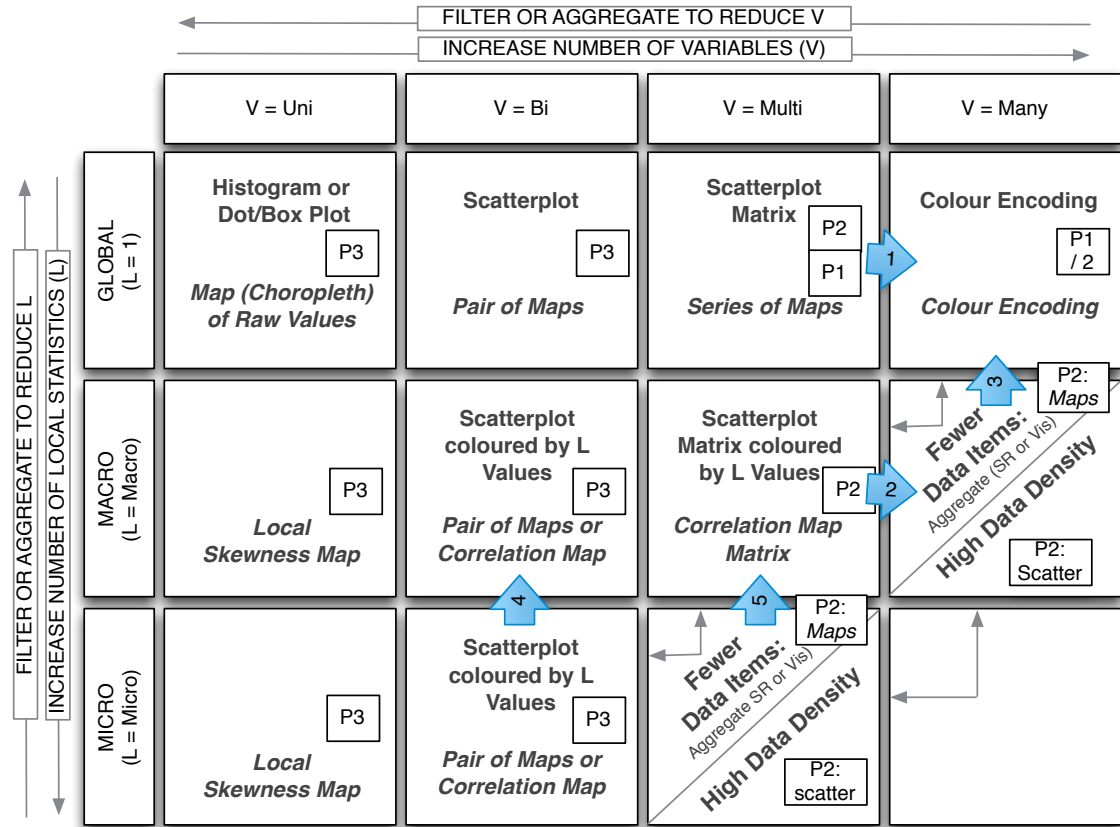


Table 6.4: The prototype illustrated within the framework with the represented statistical and geographical (italicised) visuals and panels; P1, P2 and P3. Blue arrows 1-5 highlight transitions discussed

The full breadth of the framework is not possible to represent in this prototype due to design decisions in the layout and due to the limitations of the data available. The design of the layout does not allow for all the statistical views in $V = \text{Uni}$ to be represented due to the amount of space needed for multiple histograms or dotplots to be shown at one time. Development prioritisation was placed on transitioning between $V = \text{Multi}$ and $V = \text{Many}$ as well as representing geography and scale to visualise as much information (as many parts of the framework) in one screen as possible. There are many possibilities to extend the prototype and further enhance the link to the framework, which are explained in detail in Section 7.7.

6.4 Chapter Summary

In summary, this chapter introduces a new framework for the visual representation of multiple variables across scale and geography. The framework is divided into four columns representing the number of variables (Uni, Bi, Multi or Many) and divided into three rows relating to the number of locations for which local statistics are calculated (Global, Macro and Micro). The ability to make visual comparisons within each cell of the framework is investigated and potential visual representations are proposed. The framework enables

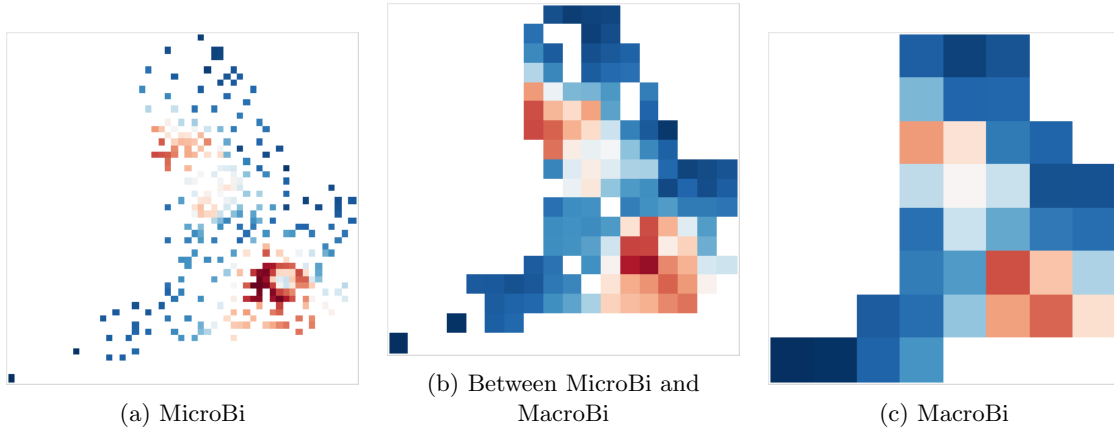


Figure 6.21: Transitioning from MicroBi to MacroBi using the local correlation map in P3 as an example of very small map squares and large ones causing the raster map to become more abstract

the comparison of many variables across scale and including geographical variation. This is a complex parameter space that is difficult to visualise without guidelines. Through the theory of scale, local statistics and the principles of visualisation, the framework has been built and potential representations investigated. A prototype to demonstrate the feasibility and effectiveness of the framework, in the context of variable selection for geodemographics, has been developed by building on these ideas.

The prototype is designed to represent the framework to allow the ability to visualise many variables across scale and with the inclusion of local geography. The complexities of scale are demonstrated through the visual comparison of four aggregations at the StR stage. For two of these four, three differing numbers of neighbours (N) for the adaptive moving window approach are demonstrated at the LR stage. The thresholds in the framework when $V = \text{Multi}$ become $V = \text{Many}$ are demonstrated through the change in visual representations and the use of spatial aggregation. The three panel layout of the prototype was implemented in order to visualise as much as possible on one screen as this was seen as important to aid viewers in their understanding of the whole concept of the framework and the parameter space. In general the design decisions of position, order, colour and interaction have been explained and justified through knowledge of the data, the context and knowledge of the visualisation domain. Additional designs are demonstrated for other types of scale; both attribute and time based data. These, including many other extensions to the prototype are explained further in the following chapter, where the utility of the framework through the use of the prototype is demonstrated.

7

Validating the Prototype Design

In this chapter, the utility of the framework and use of the prototype is investigated in detail for aiding the process of geodemographic variable selection (Scenario 1), thus addressing RQ2 (see Section refsec:rq). In Section 7.1 the appropriateness of the features, functions and design of the prototype is discussed with reference to the user stories (referred to as US#) from Table 4.1. A visual exploration of the energy variables using the prototype is carried out in Section 7.2 where the benefit of the addition of locality is investigated and some suggestions are made as to whether to include/remove variables from a classification (addressing RQ1). The investigation of the scenario continues with a visual exploration of the sensitivity of the parameters of scale, geography and transformation in Section 7.3 (addressing RQ3 and RQ4). Each are shown to be important factors when considering variables for clustering and the visualisation prototype proves to be beneficial to understanding the geographical characteristics of the variables. The results highlight that there are many complex answers to energy variable analysis demonstrated through gvPSA in this context and that a more dynamic noClassification approach to energy-based consumer profiling may be more suitable to future needs. This is discussed further in Chapter 8. Overlapping OAC variables are investigated visually in Section 7.4 and expert feedback of the prototype is reflected on in Section 8.4. The evaluation of the utility of the framework through the use of the prototype design helps to define a number of recommended extensions for implementing an improved and

deployable solution (as described in Section 7.7). The scenario is concluded by summarising the findings and new knowledge gained about energy consumption and the multivariate geography of the UK, through visualising the data using the framework.

7.1 Appropriateness of the Prototype Design

The interactive visualisation prototype, described in the previous chapter, was built in order to demonstrate the framework in the context of a real-world situation. Whilst a fully working application for variable selection for geodemographics was not possible to develop, the prototype does support a number of user stories identified in Table 4.1. Potential extensions to the prototype, relating to the other user stories and additional useful features discovered through utility in context, are discussed in Section 7.7.

The adaptive visual representation of P2 allows the visual analysis of many variables at one time (addressing User Story 2 referred to as US#02) as well as detailed investigation of the variables in P3 (US#31). The consistent use of colour (US#49) and reordering (US#10) makes it possible to identify which are heavily skewed (US#01), strongly correlated (US#06) and geographically varied (US#02). The reordering (US#10) of the global statistics (in P1) is particularly beneficial to identify extreme values in these three individual dimensions, while the default ordering by theme (US#04) is helpful as a quick overview of each variable type. The three global statistics (skewness, variance of correlation coefficient and Moran's I) were chosen to represent the three dimensions of distribution, correlation and geography in P1, yet another three re-ordering options are possible in the prototype; ordering by kurtosis value, maximum correlation and medium correlation. All seven (including theme) ordering options produce very different visual structures in P1 and P2. This allows relevant variables to be identified for further exploration (see Fig. 6.3).

As there are three dimensions (distribution, correlation and geography) to consider, it is difficult to decide which aspect to investigate first (to address US#25). The ability to quickly mask the non-extreme correlation and skewness coloured cells from the matrix (in P2 only) ensures that only data values defined as 'extreme' are visible (US#09) and aids the knowledge discovery process. The threshold of extreme is defined as ± 0.65 for correlation and ± 2 for skewness in the prototype design (as discussed in Section 6.3.2: 'Show Extreme Values'). These thresholds can be amended by the user depending on the application. A fifth column in P1 (or additional ordering option) which combines the three dimensions into a derived statistic or index would be useful to increase usability and efficiency of identifying which variables are essential for the user to investigate. Automated flagging could be used to highlight problematic variables – such extensions are discussed further in Section 7.7.

In terms of enabling the geographical variation of variables to be better understood, the prototype allows the user to see local statistics mapped (US#15) for many variables, to quickly switch between generalised (asymmetrical P2) and detailed (P3) statistical and spatial views (US#11) and enables the difference between local and global values to be displayed for each variable (US#14) through the variation in colour. In terms of multiple scales, the scale mosaic view allows for the visual representation and interactive comparison (US#30) of multivariate correlation at different SR (US#18). The transformation of the distribution of a variable is also possible to visualise in juxtaposition with the non-transformed data (US#22). Additional detail (e.g. US#19) about the variables and data items (such as values and source information) is available in the InfoBox. This textual information, along with the use of multiple views (US#50) and interaction, allows the user to relate the visuals to the underlying data (US#48)

In general, many of the user stories have been addressed through the instantiation of the framework. A number of additional user stories refer to gaining information about the variables including US#03; *“I would like to know which variables are geographically ‘interesting’, so that I can create better clusters”*, US#11; *“I would like to see variables duplicating information, so that I can decide to remove them”*, and US#17; *“I would like to know when scale has an impact on correlation so I know which variables are sensitive to scale”*. These are investigated during a thorough visual exploration of the energy variables and the parameter sensitivity described in the following sections.

7.2 Visual Exploration of Energy Variables

The process of selecting variables for clustering is investigated in this section through using the prototype. The investigation concentrates on the structure (skewness) of the seven energy domain variables and their relationship (correlation) to the other 71 variables. Firstly, the global views are explored using the reordering and ‘show extreme values’ options. This is initially investigated across all four SR levels separately for distribution, correlation and geography. Then correlation and skewness is compared across SR using the scale mosaic view. This follows with an investigation of the use of local values in the process, where local variation with and between the variables is explored. Potential conflicts between the energy variables and other variables are identified at each stage. To conclude the exploration, knowledge gained about the variables and the benefits of the addition of local values is summarised.

7.2.1 Global Skewness

Firstly, the four central heating fuel type variables are identified as being potentially too skewed for clustering, with households with other, none and electric central heating all

having a skewness value of greater than 2 for many of the SR levels, while ‘gas central heating’ has a negative skewness value of less than -2 for the three lowest SR. This is not surprising given that the overall majority of the population have gas heating (see Section 4.3). For clustering a decision would therefore be needed as to whether to keep the skewed variables, introduce weighting or implement data transformation. Although transformation options are limited to the log scale in the prototype, the log and non-log distributions can be investigated. For the first three variables, transforming to the log scale is shown to be beneficial for removing some skewness and improving the distribution, but for gas, as it is negatively skewed, the skewness worsens when transformed (as shown earlier in Fig. 6.9). The decision as to whether to keep, remove or transform the variables will depend on other features of the variables and is therefore noted for future reference. Here, flagging or annotating would be a useful feature.

It is noted that a drawback of the current design is that the ordering of P1 is ranked from high to low. While this is suitable for variance of correlation, for skewness and Moran’s I both the positively and negatively skewed variables are important to investigate. The few negatively skewed variables are removed from view when the number of variables are reduced from Many to Multi in P2. Instead, ordering by magnitude would be useful or an additional reversed ordering option. Alternatively, the *kurtosis* value (see Section 2.4.1) ordering could be used instead of skewness, as it works on histogram shape rather than direction. Ordering by kurtosis was implemented as an additional ordering option and when utilised it brings the negatively skewed ‘gas central heating’ variable back into focus. Whilst skewness was chosen for P1 for consistency with the skewness shown in the local calculations (as it was the only variable for identifying variance of distribution in the GWModel package), this highlights the need for continued testing and further development for implementing a usable application in this context.

7.2.2 Global Correlation

At the LSOA level, the noticeable strong global correlations for the seven energy variables are recorded. ‘Other central heating’ has a strong negative correlation with ‘gas central heating’ – a likely pattern as these are almost inverse variables (although electric heating is the alternative) – and also has a strong positive correlation with ‘population employed in agriculture and fishing’, which is likely to reflect the fact that gas is often not available in the sparse rural areas, where other types of heating such as oil, wood and coal are used. ‘Average electricity consumption’ is also strongly positively correlated with ‘work from home’.

As the SR increases, the number of variables which correlate with the energy variables increase. ‘Gas central heating’ strongly correlates at both the LAD and NUTS2

level with ‘electric central heating’, ‘average gas consumption’ and ‘average electricity consumption’. From the central heating variables, the ‘gas central heating’ variable would likely be dropped prior to clustering at most scales, especially as it suffers from a negative skewness and is duplicating information as ‘other central heating’ and ‘electric central heating’ cover the two alternative fuel types.

At the LAD level, other energy variables show strong correlations, in particular ‘fuel poverty’, which correlates strongly with ‘economically inactive: providing unpaid care’ (positive), ‘economically inactive: limited long term illness’ (positive) and ‘employed in ICT’ (negative), the first two of which are likely to be due to the fact that ‘fuel poverty’ is a derived variable based on identifying the vulnerable population. At this SR (and above), this variable may need to be removed or weighting introduced. Along with the previously mentioned variables, ‘average electricity consumption’ also becomes noticeably correlated at the LAD and NUTS2 levels, as it is strongly correlating with ‘detached households’ (positive), ‘lone parent households’ (negative) and ‘unemployed’ (negative). Both ‘detached households’ and ‘unemployed’ correlate with a number of the other (non-energy) variables at this level. This increased number of correlations as SR increases past LSOA level is reflected across many of the variables, indicating that the number of variables deemed suitable for clustering will reduce (or weighting or transformation becomes necessary) as SR increases, as it becomes more difficult to find non-correlating variables at these generalised scales.

7.2.3 Global Geographical Variation

The use of global Moran’s I in P1 allows variables to be ordered by the degree of spatial auto-correlation. Although limited in range (as discussed in Section 5.5.2) geographical differences can be seen between those with higher values than those with lower values when using Moran’s I in combination with the distribution maps (as shown earlier in Fig. 6.2).

In terms of identifying patterns in the energy variables, ‘fuel poverty’ has the third highest variable of Moran’s I at LAD level with a value of 0.33, indicating it is more clustered than the other variables. This is not, however, extremely clear from inspection of the distribution map, although it is evident that there is a lighter area (lower values) around London and the South East and some darker (higher values) areas in the North. Upon further inspection of the variable, while it shows a normal distribution at the LAD level, the variable at the LSOA (Input Resolution) level is positively skewed. This is expected as only a proportion of households are classed as being in fuel poverty. The distribution map at the LSOA level shows some clusters around the coastal regions and the same lighter region in London. As there is no Moran’s I calculation for the LSOA level (as explained in Section 5.4.3), no conclusion can be made on the use of Moran’s I;

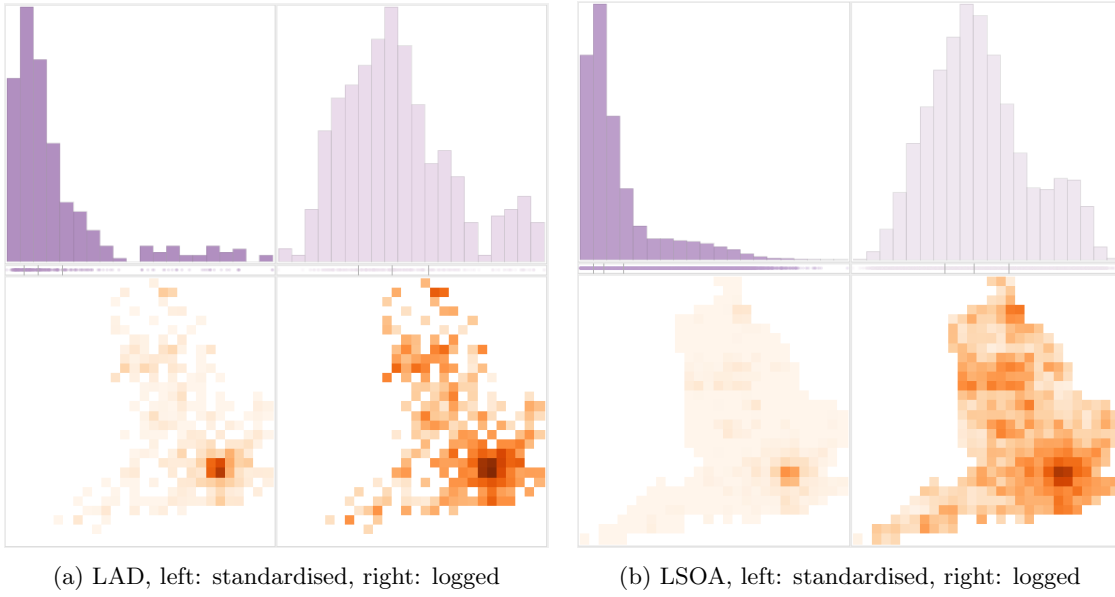


Figure 7.1: Histograms and maps showing standardised (left) and logged (right) statistical and geographical views for a. LAD and b. LSOA for the variable ‘travel to work by public transport’

however, it is evident on closer inspection of this variable that more detailed information about the survey data used to create this composite variable (see Section 4.3.5) is needed before conclusions can be made and before including the variable in the clustering process.

Using the prototype, another variable is discovered through ordering by Moran’s I; ‘travel to work on public transport’ has a Moran’s I value of 0.26 at LAD level and the map shows two distinct clusters: London and elsewhere (see Fig. 7.1a: standardised). While the Moran’s I value is relatively low compared to a value of 1 (perfect correlation), it has a relatively ‘uninteresting’ geographical pattern when visualised. This variable is strongly correlated with many other variables at this SR level and would likely be considered for removal. Although, given the nature of the case-study, those who travel to work on public transport outside of large cities could be more environmentally (and energy-use) conscious. Whilst Moran’s I was not calculated at the LSOA and lower level, when investigating the other global statistics at this level a relatively strong correlation with the other variables continues and the geographical pattern remains ‘un-interesting’ until the variable is transformed from the positive skew to the near-normal distribution using the log transformation (as shown in Fig. 7.1b: logged) and clusters of darker colours appear in the rest of the country. Further consideration of this variable is needed before a decision can be made as to whether to include it in the clustering, yet it is evident from the investigation that scale and transformation both play an important role and can influence the geographical distribution of the variable.

The inclusion of Moran’s I in the prototype allows for the variables to be ordered by spatial auto-correlation. Whilst this has helped discover some geographical patterns and

characteristics of the variables it has been the ability to visualise the detail of the variables in combination with the reordering which allowed this to happen. Moran’s I on its own, without the ability to select and investigate the variables, is not deemed particularly useful at this stage of the evaluation.

7.2.4 Global Statistics across Scale

As Moran’s I was only calculated for LAD and NUTS2, a comparison across all four scales was not possible; however, the degree of skewness and correlation is shown to be dependant on the scale of the data being investigated. The ordering of all variables in P2 by median correlation at each SR reveals a similar visual pattern of multivariate correlation at each level, yet the variable order differs in each case. Noticeably, the order of the variables is similar for the two larger SRs and the two finer resolutions, yet there are large differences when comparing the two types of generalised (NUTS2 and LAD) and neighbourhood level (LSOA and OA) aggregations. The variable ‘economically inactive: full time student’, for instance, which is at the top for NUTS2 and LAD, drops to mid-table for both LSOA and OA, indicating that the pairwise relationship of this variable is affected by scale. The fact that students cluster spatially was also noted earlier when referring to households containing only full-time students (see Fig. 6.2).

The impact of scale is also illustrated when visualising only the values classed as extreme when using the ‘show extreme values’ option. Figs. 7.2a-d show the matrix (of P2) of each SR, when only the strong correlation values are visible. It is evident that the number of strongly correlated variables increase with each SR and that the number of extreme values at the NUTS2 and LAD levels is far greater than at LSOA and OA. As the prototype contains mainly OAC variables and these variables have been chosen to be non-correlating at the OA level, the extreme correlations shown at this level in Fig. 7.2d relate only to variables which were amended for OAC 2011 or are correlating with the new energy variables. These correlations are mainly variables relating to ethnicity groups, age groups or birth location which were combined for OAC 2001, yet separated for OAC 2011 (see further variable details in Section 7.4 and Appendix B.9).

In order to visually comprehend the differences at these scales, the scale mosaic view is used with the ‘show extreme values’ option. This represents the same data as Figs. 7.2a-d combined together in one view, where patterns are easier to interpret, although the number of variables need to be reduced in order to show the multiple scales (i.e. V = Multi). Fig. 7.3a shows the scale mosaic view when the central cell is the smallest geographical unit (OA) and the outer cell is the largest (NUTS2). This view is used as default for the prototype as the largest geographical area associated with the largest square size is more intuitive. The reversal of the order so the smallest areas (OAs) are

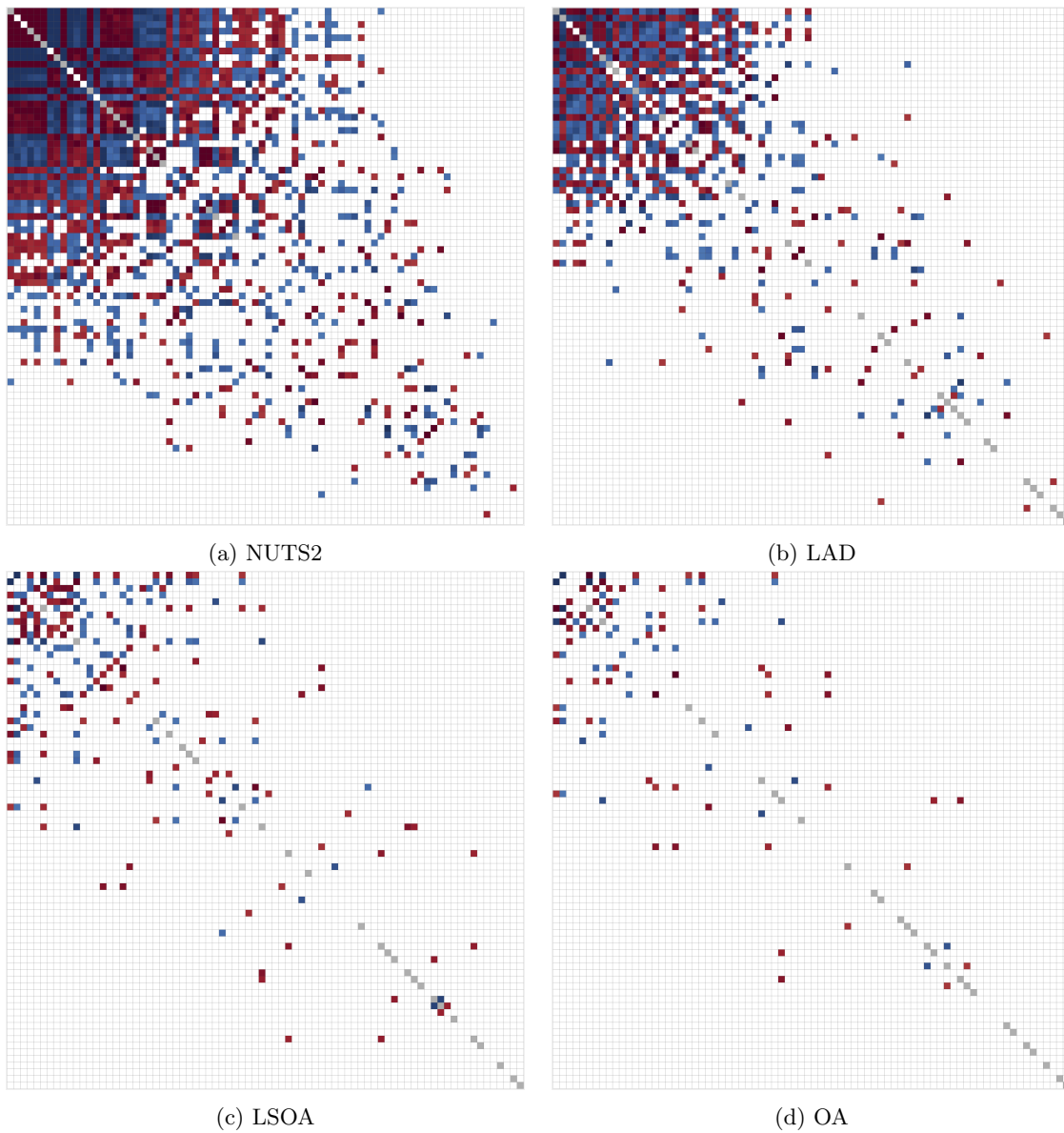


Figure 7.2: P2 showing only strong (± 0.65) correlation pairs, ordered by variance of correlation for NUTS2, LAD, LSOA and OA

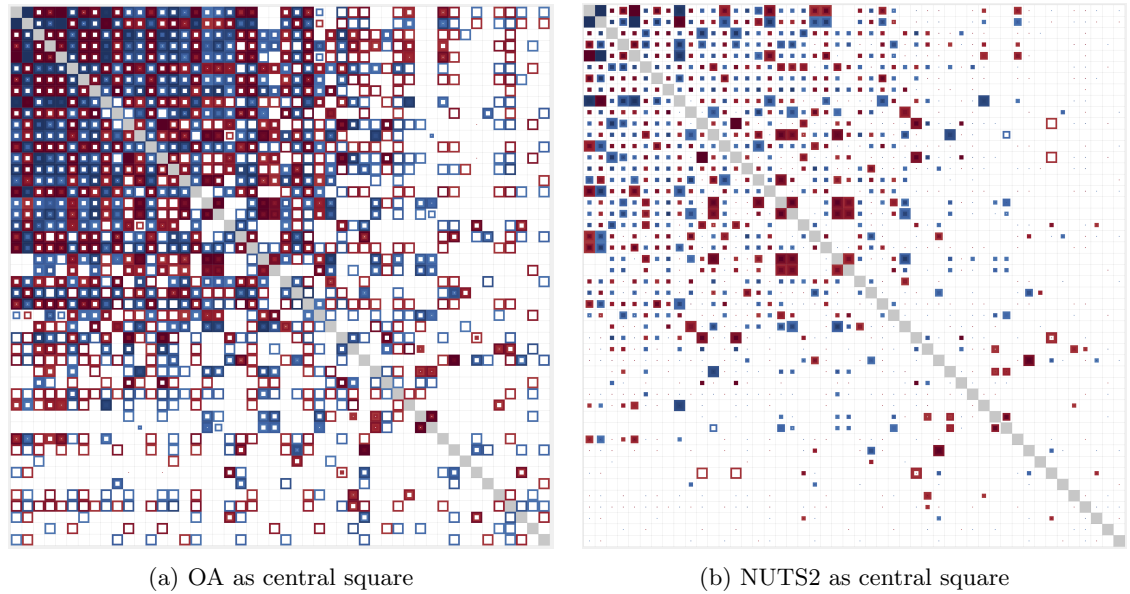


Figure 7.3: P2 showing strong (± 0.65) correlation values, for all four SR levels with smallest regions in the centre (a) and largest regions in the centre (b). The first 48 (of the 78) variables when ordered by variance of correlation at LAD level are shown

represented in the outer square of the cell is also possible (see Fig. 7.3b) and reveals some unusual geographical patterns in greater detail. As the largest geographies tend to show most extreme values, reversing the order can help to identify which variables are different to the norm; for example, a few variables *weaken* in their correlation as they increase in SR or *fluctuate* in value (see Section 7.3.2 for details). Using the ‘show extreme values’ option can therefore not only help to identify variables, but can begin to identify which variables are sensitive to aggregation at these scales. Scale is explored through sensitivity analysis in Section 7.3.2. While it is evident that the visualisation of the global data as a dynamic and interactive matrix (P2) helps to identify which variables are relevant to investigate further for variable selection, additional automated processes could benefit the creator (user of the tool) and improve the usability of the prototype.

7.2.5 Adding Locality into Variable Selection

The inclusion of locality in the variable selection process is of particular interest in the context of this scenario as local statistical variation in variables is currently not reported as being included in the process of generating geodemographic classification. The asymmetrical matrix is beneficial to help investigate the local variations across the geographical and statistical space (as shown in the video¹ at 2m:40s and in the examples below). Many variables are identified through using the asymmetrical matrix, in particular through comparing the raster maps. Fig. 7.4 shows the asymmetrical matrix for all seven variables in the energy domain. The background of the scatterplot view

¹this is available in high quality in the digital appendix submitted with the thesis and online at: <http://vimeo.com/112182748>

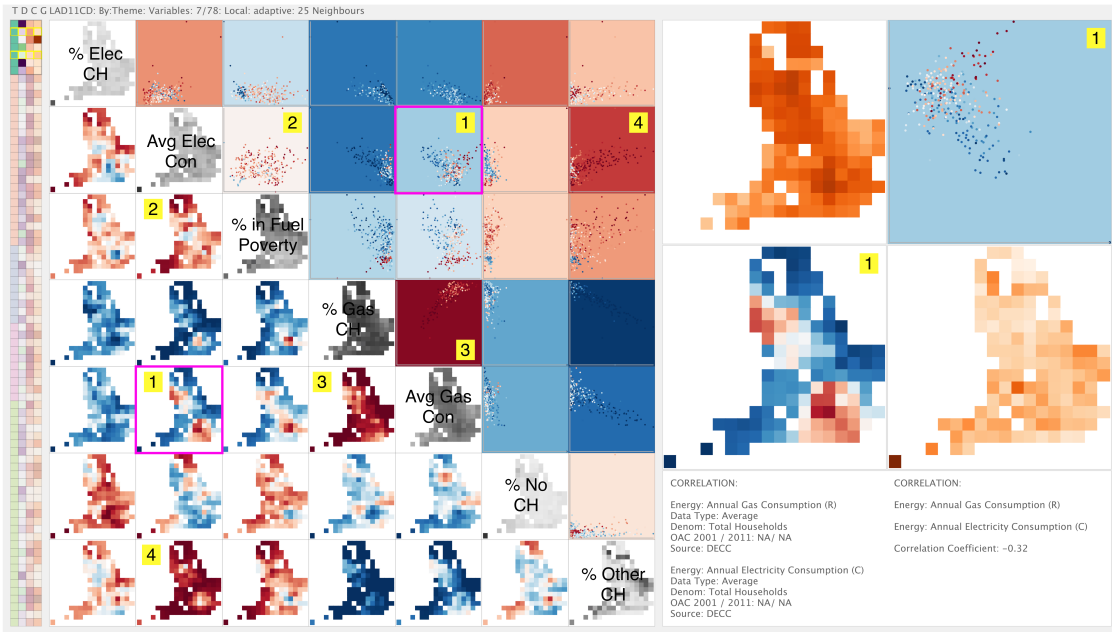


Figure 7.4: The prototype showing all seven energy variables with locality (adaptive moving window with 25 neighbours) information shown in the spatial and statistical view. The colour of the scatterplot background represents the global correlation coefficient. Four examples are highlighted and numbered.

represents the global correlation coefficient to quickly distinguish variables that have a variety of different local values in comparison to the global value.

Example 1 (see 1 in Fig. 7.4) compares electricity and gas consumption and is shown in P3 as well as P2. This example displays interesting patterns at the local level with a range of local correlation values from strongly positive in London to strongly negative in rural areas (this example was also discussed in Section 5.1). Example 2 (see 2 in Fig. 7.4) identifies a pair of variables where the global correlation coefficient shows that the variable has no correlation at all; ‘Average electricity consumption’ and ‘fuel poverty’ has a global correlation coefficient of 0.03 at the LAD level (LSOA has a global correlation of 0.16), yet the local map and scatterplot shows there is in fact correlation between these two variables across most of the England and this ranges from very strongly positively correlated (above 0.8) to highly negatively (under -0.5) in some areas. Examples 3 and 4 (see 3 in Fig. 7.4) are similar. They both show strongly positively correlated variables at the global level, yet these range across the country with ‘average gas consumption’ and ‘gas central heating’ (Example 3) with many paler data points in the centre of the country and ‘average electricity consumption’ and ‘other central heating’ (Example 4) with some negative (blue data points visible) correlation.

The visualisation of local skewness is also interesting and helps reveal that some variables deemed as extremely skewed are in fact not skewed across the whole country. ‘Electric’ and ‘gas central heating’ are both classed as extremely skewed globally but the skewness pattern across the country differs. Notably, nearly all the local maps in the

prototype indicate that local patterns are different in the densely populated urban areas, with London and the North West of England often appearing as extremely different from the rest of the country. This outcome is reduced as the number of N (neighbours) increases and the values converge to the mean of the full extent (StE) of the dataset under consideration (as shown in Fig. 5.2 and investigated in the following section). This additional local information is likely to be very useful to identify variables, which may be affected by a change in SE, if multiple classifications are created using the same or similar variables. For instance, variables with significantly different local correlations in London may be useful for a nationwide profile, but may not be suitable for use when the SE is reduced to London.

7.2.6 Summary of Visual Exploration of Energy Variables

This section demonstrates the use of the prototype for variable selection and the ability to investigate the variables in depth. Interesting patterns and fundamental differences are found when visualising the local statistics instead of just the global values. The ability to transition between Global, Macro and Micro is useful for progressing from identifying strongly correlating variables at the global level and subsequently investigating and better understanding their geographical variation at the local level. Although suggestions have been made on the use of certain variables (e.g. ‘gas central heating’ to be removed), further investigation is needed in order to interpret exactly which variables would create a good classification for the energy industry as this will involve the running of clusters to test the impact of the variables and this is beyond the scope of this study. However, it is evident from this inspection and reference to the literature that additional Census variables which do not appear in OAC 2001 or 2011 would be useful to combine in this variable selection process relating in particular to household composition. Family types are not included in OAC as they correlated with age groups and ‘in full-time education’, yet these would be potentially more useful in an energy-based variable than their alternatives. While more time and research is needed in order to test a larger number of socio-economic, household and demographic variables for relevance to energy, it is evident that the prototype is useful in this context, allowing the dimensions of variables to be visualised and decisions to be made based on variable distribution, correlation and geographical variation. The visual exploration of the energy variables continues in the following section, which investigates the sensitivity of the parameters.

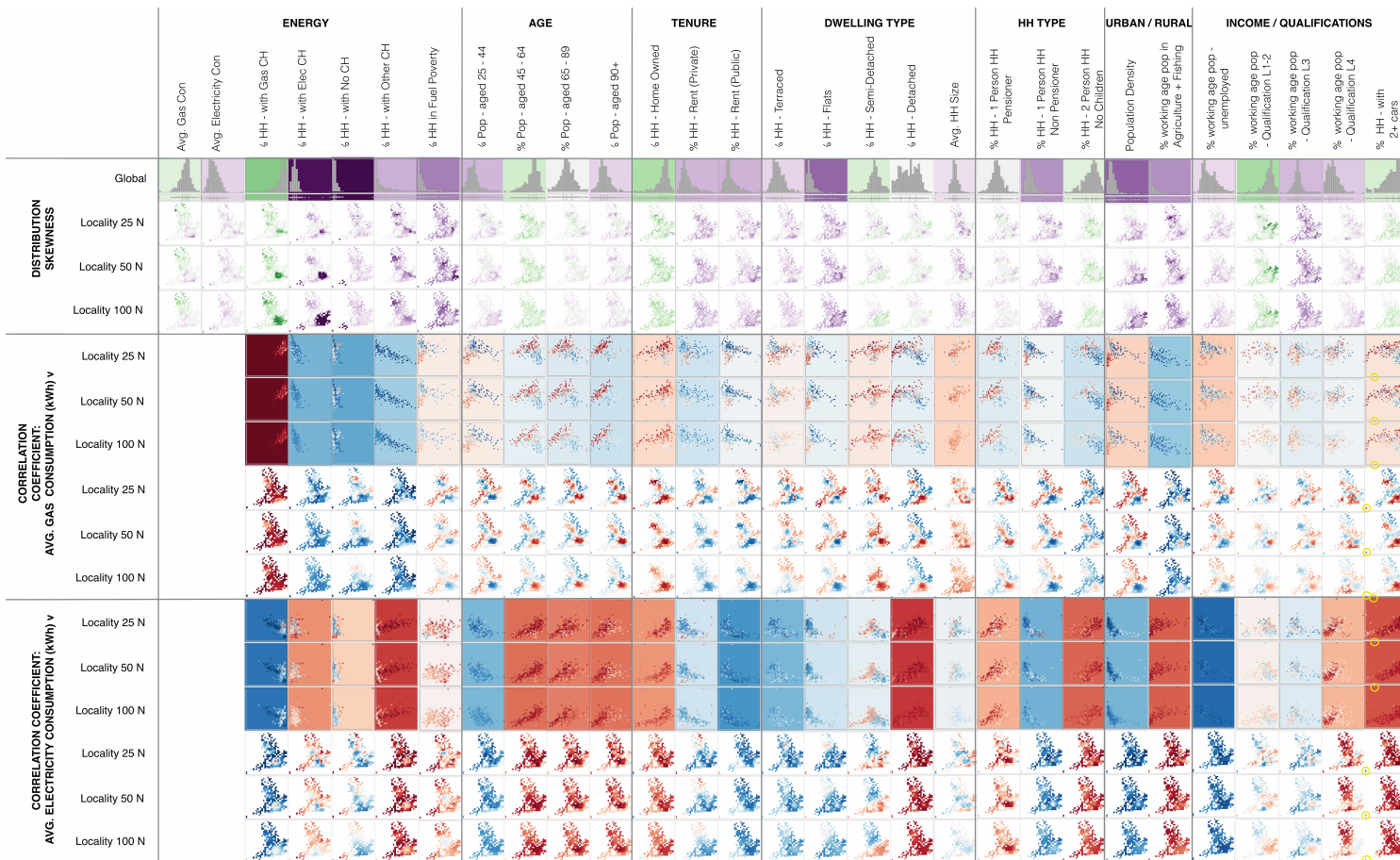
7.3 Geo-visual Parameter Sensitivity Analysis

In order to answer RQ4, the sensitivity of varying geography and scale (and to some degree transformation) is investigated. This has partly been investigated through the

analysis of the variables in the previous section. A more in-depth visual investigation is necessary in order to understand the effect of varying scale, geography and transformation on the visual representations. For this, it is noted that the prototype can only represent a proportion of the variables in P2 before the pixels run out and the global value is used. In order to visualise the local (e.g. Micro and Macro) detail of multiple variables, the ‘more pixels’ option of MacroMany or MicroMulti is investigated and additional graphics are used for this section, utilising the graphical components of the prototype (the grid-like maps, scatterplots, histograms) with the same data. This enables the visual evaluation of all the graphical representations with all the dimensions of the variables in juxtaposition, rather than restricting the visuals to those designed to be compared within the prototype. Certain variables are noted as being sensitive to varying geography, scale or data transformation.

7.3.1 Sensitivity of Geography: Varying N

As geodemographics are inherently geographical (see Section 2.4.3), variation in geography and sensitivity to locality are important to consider in variable selection. Sensitivity of geography is investigated in this section through exploring the visual representation of varying the neighbourhood parameter (N) in the adaptive moving window approach to locality calculation. This analysis uses the two SR levels of LAD and NUTS2, with three differing LR: $N = 25, 50$ and 100 for LAD, and $N = 5, 10$ and 15 for NUTS2. An in-depth investigation of which locality method and parameters are optimal for the context is beyond the scope of this research study, yet the initial investigation outlined in this section demonstrates that the scale of the data and the value of N can greatly affect the output results.



7.3.1.1 Local Skewness

Table 7.1 shows the graphical components of the prototype compared in juxtaposition for a number of the key variables for the most detailed local data available (LAD). The table shows variable skewness and the variability of correlation in relation to average gas and electricity consumption, as these are key to the case-study. These two variables also differ in their skewness and distribution, as shown in columns 1 and 2. Although both show geographical variation, there are areas of no or limited gas and therefore it is expected that local variation is particularly evident. The variables in the table are chosen from suggestions in the literature (Druckman and Jackson, 2008; McLoughlin et al., 2012) that energy use is heavily related to the age of the residents, household tenure, dwelling type, household composition, urban/rural location and household income (as discussed in Section 2.1). ‘Household income’ is not available in the Census (and not included in OAC), so alternative variables from the 78 seen as potential proxy variables for income were chosen. The variables relating to ethnicity and types of employment are not included in Table 7.1 as they are reported in the literature (Druckman and Jackson, 2008; McLoughlin et al., 2012, e.g.) to have less influence on energy use, although the effects of varying N are similar to those identified.

At the top of the table, the variable distributions are shown through global distribution histograms, with three local maps showing the difference in skewness as the calculation of locality changes for $N = 25, 50$ and 100 . It is evident that some variables are more sensitive to changes in N than others. The variables that are relatively normal appear to have less intense variability in the skewness maps than those which are heavily skewed. This is not surprising as outliers can have a great affect on local values. It is also noted that this can be partly due to the colour scheme, as those which are heavily skewed are projected more vividly than those that fluctuate around normal. The use of difference maps would potentially be more beneficial in this context to represent the magnitude of the differences. Difference maps are discussed as a useful and necessary extension to the prototype in Section 7.7.

It is noted by examining the graphics in Table 7.1 that variables which are heavily skewed, such as the central heating fuel types (gas, electricity and none), show more intense skewness as N increases. In these maps, relatively few highly skewed values are evident at the very local level ($N = 25$), indicating that there are only a few highly skewed values that are affecting the overall global value. For ‘gas’ and ‘electricity central heating’ it is clear that the values around London are particularly influencing the global skewness value, therefore care should be taken if the SE is reduced to London. ‘No central heating’ shows a similar effect, but due to high skewness in the very rural areas on the tip of Cornwall and the Isles of Scilly. This is an interesting observation, as

the Isles of Scilly are noted as being especially problematic in the analysis in general, particularly in reference to correlation (see yellow circles in the last column in Table 7.1) with ‘average electricity consumption’. The recording for ‘average electricity consumption’ is particularly high, potentially due to the lack of central heating or due to a change in the electricity measurement process (e.g combined meters). Further research is needed for investigating this isolated region. Nevertheless, the dual statistical and geographical views in the asymmetrical comparison matrix with the BiMacro and BiMicro in the detail panel enables these geographical outliers to be identified quickly. The ability to remove outliers from the comparison or add a threshold, would be useful extensions to the prototype and are discussed in Section 7.7.

The opposite skew effect is evident on ‘other central heating’, where there are a number of skewed values evident at the $N = 25$ level, which disappear as N increases upwards to the global level. This effect is also evident on the variable ‘employed in agriculture and fishing’. Both these variables are very dispersed geographically (via the variable distribution maps) and only appear in very rural areas. The effect of zero or near zero in most regions is causing the heavy positive skewness of the distribution locally. In the rural areas high values appear resulting in a negative skewness and therefore when aggregated the global skewness of the variable is smoothed. Again, these variables will be particularly affected by filtering the SE to urban or rural areas.

Another interesting variable from Table 7.1 is ‘fuel poverty’. This variable shows a different pattern to the others when varying N . There are limited heavily skewed values in $N = 25$, while heavily skewed values do appear at $N = 50$, yet disappear again at $N = 100$. Further research is needed to understand why this occurs only for this variable. It is perhaps an indicator that a more optimal level for N is needed for such variables, rather than choosing sequential steps for N as carried out in this study. A more optimal N is likely to vary spatially as well as across scale. Importantly the exploratory graphics, outlined in this section, enable the user to see the effect of varying N – in multiple variables and places – concurrently.

These effects of local skewness can be seen in the other non-energy variables in Table 7.1, but to a lesser degree as most are less skewed. In general the other variables are less skewed than the energy variables. Some of the variables which do not appear in the table are also heavily skewed, such as the ethnicity variables, where most variables are heavily positively skewed in comparison to ‘white British’, which is heavily negatively skewed. Similar local patterns, as noted above for the energy variables, are also evident when N is varied. Through the investigation of the skewness of the energy variables, it is evident that energy use is extremely complex to model and consider

spatially. This is likely to be because of the various spatially constrained options through which energy is consumed.

7.3.1.2 Local Correlation

When the correlation maps and scatterplots are displayed in juxtaposition (Table 7.1) and interpreted with knowledge of the geography of England, three types of geographical correlation patterns are noted: those largely urban/rural correlations (e.g. ‘gas consumption’ and ‘home owned’), those with little geographical difference across the country (e.g. ‘electricity consumption’ and ‘home owned’) and those with a North-South divide (e.g. ‘electricity consumption’ and ‘semi-detached housing’ or ‘private-rented housing’). These patterns may reflect cultural differences that are important considerations for energy-based geodemographics.

Nearly all variables when correlating with gas consumption reflect the same urban/rural pattern shown in the distribution of ‘gas central heating’ (and in ‘average gas consumption’ but this is less obvious). This is unsurprising considering gas consumption occurs only in areas where gas is available. The local correlation usually shows a reversal of the correlation in the gas areas compared to the non-gas areas, with some variable correlations being more intense than others. While in general none of the chosen non-energy variables have a strong global correlation with gas consumption, the local correlation statistics show that this is most often not the case across the extent of the study area. The local scatterplots illustrate this also clearly with a large variety of correlations ranging from strongly positive through to strongly negative. Only ‘average household size’ and all the ‘central heating’ variables show a different pattern and have less variation in hue in all levels of N.

In comparison to gas, ‘average electricity consumption’ has a far more varied relationship with the other non-energy variables (see Table 7.1) in terms of the local geographical patterns as well as the variance of the global correlation. Many variables have a strong or relatively high (positive or negative) global correlation with ‘average electricity consumption’. Only a few variables have little correlation at the global scale; notably ‘fuel poverty’, ‘average household size’, ‘semi-detached housing’ and ‘highest qualification L1-2’. In these four examples a variety of correlations are evident at $N = 25$, but these correlations disappear as N increases. Differences at the $N = 25$ level is also clear when the correlation is particularly strong, such as with ‘unemployed’ and ‘detached housing’. This differs from the pattern shown from the two strongly correlated energy variables of ‘other’ and ‘gas central heating’, where there is a large amount of variation between values in all three levels of N. Perhaps this is due to the more varied or clearly different geographical distribution of these two variables. Many of the non-energy variables do not

show significant local variation, but some are notably varied. These include all three ‘qualification’ variables, the two ‘1 person household’ variables and ‘rented privately’. Some of these variations are not visible in the raster map views, which is a drawback of the design and the spatial aggregation used (Section 6.3.1.5); however, there is a benefit from using both views and having the ability to identify items through interaction.

7.3.1.3 Scale Resolution of Neighbourhood Parameterisation Analysis

The investigation of the data in this section uses the SR level of LAD with 326 values and the LR of $N = 25, 50$ and 100 . These sample sizes were chosen to adequately demonstrate the effect of varying N , without causing the correlation results to be unreliable through small sample sizes (see Section 2.4.2). Although the NUTS2 level of data has been calculated there are far fewer regions (30 in total) and calculated local statistics have only sample sizes of 5, 10 and 15. The NUTS2 data was not analysed in the same way as LAD, as correlation results may be unreliable (Martin, 1978) and the design of the raster maps in the prototype make visualising maps with so few data points difficult to interpret. The LSOA and OA levels were also not possible to run for the research, although these would be very useful to investigate as they are based on population size and therefore the number of N will reflect the underlying population. The global correlation relationships between the variables are also much stronger at the NUTS2 level (as shown in Fig. 7.2) and this adds complexity to the comparison. Additional SR, with more data items than NUTS2, would be useful for comparing the local as well as the global data across scales. The effect of varying SR on the four global statistics is investigated in the following section.

7.3.1.4 Sensitivity of Geography Summary

In summary, there is much more information in the visual investigation of the local skewness and correlation of these variables compared to the use of global statistics. In the examples expressed, the magnitude of the variation in local skewness and correlation seems to be dependant on the variables themselves, their statistical and geographical distribution as well as the value of N . Variables that are geographically varied show greater variation in correlation at the local level and heavily skewed variables have more local skewness variability than variables that are relatively normal. Further research could involve more in-depth analysis of adapting the type of locality calculation, as well as varying N and perhaps whether a more optimal value for N can be determined. Perhaps this value is different for each variable and the visualisation needs to be adaptable. Additional functions are available in the GWModel R package for the selection of N (known as bandwidth) and these could be analysed for the initial

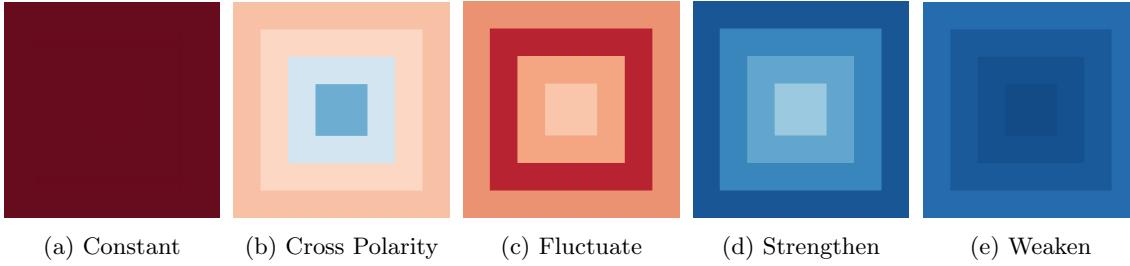


Figure 7.5: Five categories of scale sensitivity shown using the scale mosaic design. Ordered by geographical hierarchy with OA as the central cell. Colour refers to the degree of positive (red) or negative (blue) correlation

investigation, along with the comparison of adaptive and fixed (known as kernel in GWModel) locality calculation options.

Further investigation is also needed as to what extent local variation effects the clustered output. Do variables with strong global correlation, yet limited local variation, differ in their effect on the clustering from those with strong global correlation and strong local variation? While it is evident that much research is still needed, the results of this analysis indicate that it seems appropriate that some form of local statistics are considered as part of the variable selection process for geodemographics, especially considering selected variables are expected to be geographically varied and this is shown to have local effects.

7.3.2 Sensitivity of Scale: Varying SR

The effect of varying the spatial aggregation of SR on the global skewness and correlation values is visualised in the prototype using the scale mosaic view. Five different visual characteristics are noted, relating to five different effects on the values as the SR increases:

- Constant
- Crosses Polarity
- Fluctuate
- Strengthen
- Weaken

These five categories can each be identified in the scale mosaic tiles as shown in Fig. 7.5. *Constant* groups together variables which have near perfect (positive or negative) correlations and when aggregated these values do not change, i.e. there is no effect of aggregation. This category distinguishes the overlapping or inverse variables changed between OAC 2001 and OAC 2011. ‘Aged 65+’ was used in OAC 2001, whilst ‘aged 65-89’ was used in OAC 2011 (shown in Fig. 7.5a. Whilst ‘born outside the UK’

was used in OAC 2001 but removed from OAC 2011, it remains *constantly* correlated with ‘households with no English language’ used in OAC 2011. Some variables are only very slightly affected by scale but these examples are characterised into one of the other four categories.

Typically, the variable correlation and skewness *Strengthens* (Fig. 7.5d) as the SR increases and the outliers are removed; for example the correlation of the two variables ‘average electricity consumption’ and ‘gas central heating’ strengthens from -0.5 at OA, -0.54 at LSOA, -0.71 at LAD and -0.78 at NUTS2. *Crosses Polarity* is also a typical pattern for the variables which have a low correlation and upon aggregation, the value shifts from below-to-above or from above-to-below zero. When the colours within the cells of the mosaic are notably dark then the variable pair is particularly sensitive to aggregation; for example Fig. 7.5b represents the two variables ‘home owned’ and ‘separated/divorced’, which correlate at -0.48 at OA, -0.47 at LSOA, 0.04 at LAD and 0.42 at NUTS2. *Fluctuates* represents values which do not increase or decrease smooth and continuously, but vary between stronger-weaker-stronger or weaker-stronger-weaker. Fig 7.5c shows ‘detached households’ v ‘semi-detached households’, which correlates at -0.14 at OA, -0.07 at LSOA, 0.2 at LAD but 0.0 at NUTS2. These examples are particularly sensitive to changes in scale.

The two latter categories – *Crosses Polarity* and *Fluctuates* – are useful to highlight as a drawback to the current functionality of the prototype. The ‘show extreme values’ option, which is shown to be useful to indicate which variables are strongly correlated, can mask these variables which are sensitive to scale when used with the scale mosaic view. These examples are more visually apparent when the global variance of the four SR is used in the GlobalMany view, as shown in Fig. 6.17. This indicates that some additional functionality and redesign is needed for ensuring that all types of variables sensitive to scale are identifiable at each of the stages, particularly GlobalMany and GlobalMulti.

The final *Weakens* category has only a few examples in the candidate variable dataset (with a very small difference range in comparison with *strengthens*) and therefore is deemed not a typical pattern in this instance; Fig. 7.5e shows ‘average house size’ with ‘single status’, where the correlation values range from -0.82 at OA to -0.70 at NUTS2. These variables were difficult to distinguish at first, but were found by showing only the ‘extreme values’, reversing the central SR (as shown in Fig 7.3) and looking for abnormal patterns.

In general, the scale mosaics reveal that variables at OA and LSOA level have less strong correlations compared to the two more generalised SR levels of LAD and NUTS2. There is not only a large gap in the number of regions between LSOA and LAD (see Table 5.1) but the geographical regions are built for different purposes. OA and LSOA are statistical output regions, designed to relate to the population and households, while

7.3. GEO-VISUAL PARAMETER SENSITIVITY ANALYSIS

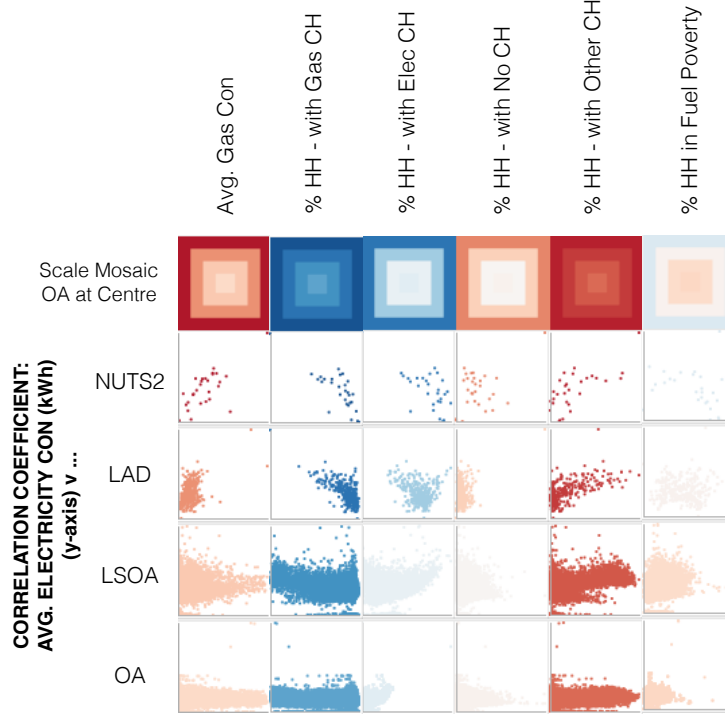


Figure 7.6: Scatterplots with scale mosaic tiles for all energy variables in relation to average energy consumption in each of the four SR

LAD and NUTS2 are administrative boundaries used for many purposes. Therefore it is perhaps more appropriate to investigate the changes between the two neighbourhood levels and the two generalised levels separately. For instance, variables which change a significant degree between OA and LSOA could be classed as very sensitive and therefore should not be aggregated. In order to investigate this effect further, additional SR would be useful to include in the geographical hierarchy, for instance the inclusion of MSOA, the third tier Census geography. MSOA was not chosen for investigation in this analysis (as explained in Section 5.4.2), but this observation indicates that there is a reason to investigate the variation at three scales (OA, LSOA, MSOA) that are meaningful and continue this investigation of the effects of varying scale.

In addition to the scale mosaic view, the scatterplots at each level of SR can help to demonstrate the actual changes to the variable structure which have occurred during the aggregation. Fig. 7.6 shows the energy variables in relation to ‘average electricity consumption’ (y-axis). The scale mosaic view combined with each scatterplot in juxtaposition is useful to see how the shape and slope of the variable relationship changes with aggregation². The scatterplots also indicate whether the variables are skewed; the effect of transformation on the visual representations is investigated in the following section.

²Note that for these energy variables the original input resolution of the data is LSOA and the OA level is estimated (see Section 5.4.1)

7.3.3 Sensitivity of Transformation: Logged Scale

In the prototype design, the ability to explore the sensitivity of transformation is limited to visualising the standardised or logged values through an additional histogram, and distribution or skewness map in the $V = Bi$ representation in P3. Therefore, as with the investigation of the sensitivity of geography, the graphical components of the prototype are displayed in juxtaposition to visualise the effects of (the log) transformation in all the statistical and spatial views (see Table 7.2). The transformation of the data is more evident in the statistical views (the histograms and scatterplots), than in the geographical views; however, the combination of the two are useful to determine whether the transformation affects a particular geographical region that may have been affecting skewness. For instance, the ‘electricity’ and ‘gas central heating’ variables have differing skewness, yet the map reveals that when transformed, the skewness in London improves for ‘electricity’ but worsens for ‘gas’. This indicates that the local map views are complimentary to the statistical views for the understanding of the effects of transformation on variable distribution.

While the scatterplots show clear differences, the global and local correlation values (encoded by colour) only show slight differences between the logged and non-logged examples. In this example this is because both variables are logged and therefore the correlation between two variables is only affected when one variable is affected by the transformation more acutely than another, i.e. ‘other central heating’ shows a slight change in the correlation map when logged, with less extreme negative values and some more positive values.

This investigation mirrors knowledge on the effects of logarithmic scale on different types of variables and further investigation is needed for other transformation options, such as the other two tested for OAC 2011: Box Cox and Inverse Hyperbolic Sine (Gale, 2014b). Although only one type of transformation was investigated and minimally included in the prototype, it is clear from this brief investigation of the visuals in Table 7.2 that a visualisation system with the inclusion of local and global statistics with multiple views is beneficial in the understanding of transformation on multiple variables.

7.3.4 Summary of Parameter Sensitivity

The parameter sensitivity investigation in this section reveals that many variables are affected by varying geography, scale and transformation, yet some far more intensely than others. In terms of local skewness, some variables are more sensitive than others to changes in N (neighbours) in the locality calculation. Those with relatively normal distributions seem to have less variability in local skewness than those that are heavily skewed. The degree of variation of local correlation also seems to be dependent on the geographical

7.3. GEO-VISUAL PARAMETER SENSITIVITY ANALYSIS

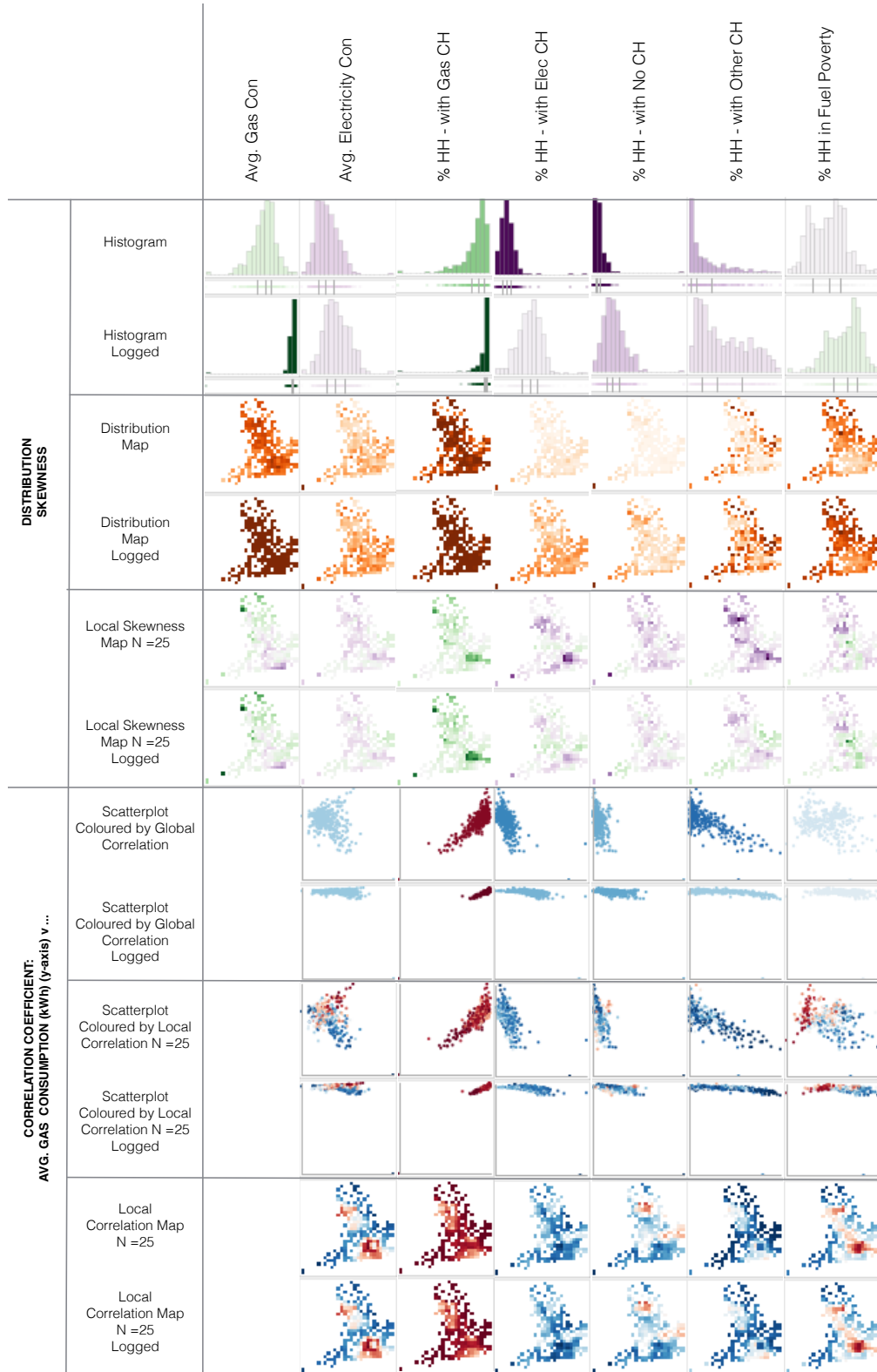


Table 7.2: The effect of transformation for the energy variables, shown in histograms, distribution maps, local skewness maps, scatterplots (with global and local encoding) and local correlation maps

variation of the variable as well as the number of N (neighbours), with variables with geographically varied distributions also showing geographically varied local patterns. For scale, it is evident that correlation generally gets stronger (strengthens) with aggregation (SR), yet four other alternative types of response are demonstrated through the scale mosaic view, with some variables affected very differently than others. The investigation of transformation reiterates that some variables are more sensitive than others to the log transformation. While further research is needed to understand the extent to which these dimensions affect the clustered output, it is evident that this instantiation of the framework is useful for exploring and beginning to better comprehend these sensitivities on different variables as well as showing variability, detail and nuance. It is also evident that when selecting variables for geodemographics the dimensions of scale and geographical variation should be considered. As shown in this exploration, there is not to a simple answer but large and complex answers where visualisation designs based upon the framework can be beneficial to understanding some of the complexity.

7.4 Identifying Discriminating OAC Variables

There are a number of variables in the 78 chosen to be in the demonstrative dataset (Section 4.3) that were adapted between the two classifications of OAC 2001 and 2011 (further details of Census table sources in Appendix B.9). This section describes how the prototype is used to investigate whether the new OAC 2011 variables are more geographically discriminating than the 2001 alternatives. OAC was produced from OA data and while the global correlation is used for a basic overview, the local statistics are not available at OA level, so LAD is used to demonstrate the geographical variation in this instance. While an in-depth investigation is beyond the scope of this thesis, a preliminary visual exploration and knowledge of the variables reveals some interesting patterns.

‘Born outside the UK’ (OAC 2001) was removed for OAC 2011 and replaced with the exact inverse ‘born in the UK’ together with two additional variables: ‘born in the old EU (pre 2004)’ and ‘born in the new EU (post 2004)’. ‘No English language’ was also added to 2011. Age attributes were amended slightly with ‘aged 65+’ (OAC 2001) replaced with ‘aged 65-89’ and ‘aged 90+’. Additionally the attribute scale extent was increased in the case of ethnic group variables, with the variable ‘white’ added for 2011 having been omitted in 2001. Attribute scale resolution also increased with respect to ethnic group, where variables grouped together for 2001 were separated for 2011 (see Appendix B.9 for details).

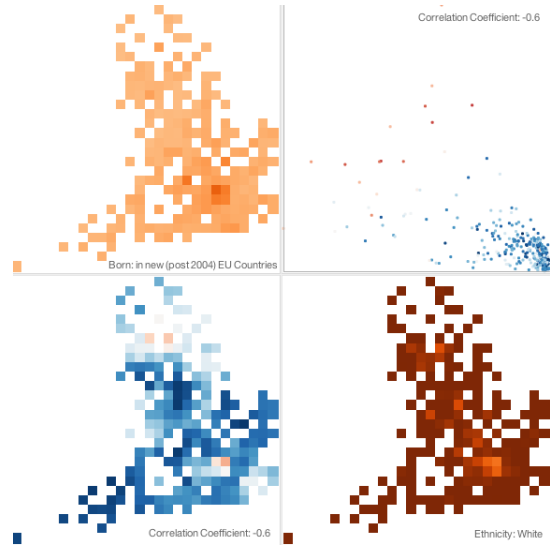
The OA GlobalMany view (see Fig. 7.2d) and the scale mosaic GlobalMulti views with the ‘show extreme variables’ option (e.g. see Fig. 7.3b) were first used to discover the

global correlation across all four SR for these similar or inverse variables. With the exact inverse variables showing the *constant* sensitivity pattern as discussed in Section 7.3.2, and many of the others showing the most common *strengthen* pattern. Visual exploration of the scale mosaic shows that ‘born outside the UK’ has a strongly negative correlation with ‘white’ and the pattern is, as anticipated, the exact inverse for the ‘born in the UK’ variable. The strength of these positive and negative correlations are strong at both OA level data returning 0.87 (or -0.87) and 0.94 (or -0.94) for NUTS2. The two ‘born in the EU’ variables show limited global correlation with -0.35 (new EU) and -0.4 (old EU) to ‘white’ at the OA level, with this increasing with scale to -0.75 (new) or -0.77 (old) at the NUTS2 level. The inversely associated variables ‘born outside the UK’ (e.g. see Figs. 7.7c) and ‘born in the UK’ have strong correlations with many of the other variables in OAC and show little geographic variation in this strong correlation across LADs. This investigation shows that whilst OAC 2011 still contains a variable (‘born in the UK’) that correlates highly with many other variables, the two EU variables add some discriminating benefit and have more varied geographical patterns at the local level (for LAD, e.g. see Figs. 7.7a and 7.7b). Although the OA level locality data has not been generated, the scale mosaic shows the global values are far lower for OA than LAD. Further research is needed in order to explore their geographical variation at the OA level.

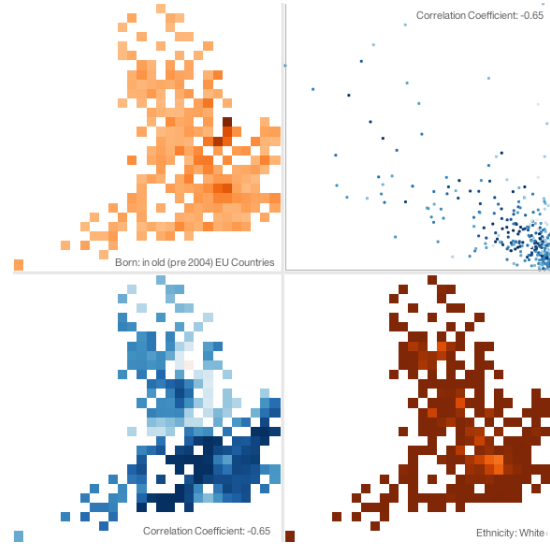
This exploration demonstrates that the use of the inverse variable (‘born in the UK’) and addition of the two new EU variables allows the new profiles to be more discriminative to the cultural backgrounds of the population. This was indeed a goal of OAC 2011 as OAC 2001 had been criticised for the Super Group ‘Multi-Cultural’ dominating urban areas, in particular in much of Greater London (Gale and Longley, 2012; Gale et al., 2014; Gale, 2014b), as discussed in Section 2.3. It can be noted that whilst these cultural or ethnic distinctions are seen to be important for general population profiles such as OAC, they are less likely to be essential for energy profiling, where housing, family size, economy and geographical location play an important role. Even if relationships can be determined, care must also be taken when choosing particular variables to ensure that they are appropriate to the chosen domain (as discussed in Section 2.3). Whilst a more in-depth investigation of these OAC variables is beyond the scope of this thesis, the observations noted in this section demonstrate the advantage of visual interactive exploratory tools for selecting variables and allowing for the visual distinction of the effects of geography and scale.

7.5 OAC Expert Feedback

An informal feedback session for the prototype took place in September 2014 (at the giScience conference) with Chris Gale, who created the 2011 OAC (Gale, 2014b). The



(a) 'white' and 'born in the new EU (post 2004)'



(b) 'white' and 'born in the old EU (pre 2004)'



(c) 'white' and 'born outside the UK'

Figure 7.7: MicroBi view in P3 of 'white' (new in OAC 2011) compared to the two 'born in the EU' variables (new in OAC 2011) both showing reduced global correlation compared to the previous 'born outside the UK' (removed from OAC 2011) variable and a more discriminating geographical variation in local correlation maps at the LAD level

initial re-ordering of the interactive grid of global correlations was seen as a huge advantage to the static view which he had used. The ability to quickly dive deep into the data was also seen as very useful, in particular for investigating and comparing the geographical variation of variables. This was carried out in the 2011 OAC by investigating a number of urban areas and seeing how the profile varied (as discussed in Section 2.4.3) and was noted as difficult to do for many variables. Usability of the tool was discussed and it was seen as too complex for variable selection for geodemographics at present and some filtering of features would be needed to allow and improve the efficiency of use. The ability to view multiple scales through the scale mosaic view was seen as less important in the case of OAC 2011, as the scale had already been decided prior to the creation. However, it was seen as very interesting and useful for deciding on variables for local or domain-specific geodemographic creation and to allow the comparison of Census variables together with other open data sources at different scales.

7.6 Visualisation for Geodemographic Variable Selection

The detailed exploration of the energy variable and varying the parameters revealing sensitivities associated with data scale, geography and transformation illustrate the importance of these factors in multivariate geographical comparison. Having explored the variables in this chapter it is now known that those relating to energy (and most of the others) vary with scale and geography and that they are variably sensitive to these variations. This suggests that the incorporation of data at different SR and for different SE within the same classification may help to represent the social phenomena which operate at different scales and at different geographies. While it is evident from this investigation that the information gained from visualising the local patterns is informative and improves the understanding of the variable pair relationship, it is unclear as to exactly how beneficial this additional information is to the classification itself. The evaluation of the impact of how variables with geographically diverse correlation or skewness affect the clustering algorithm is beyond the scope of the project, but poses useful questions for future research. An investigation of how selected input variables affect the final classification output would also increase the link to vPSA research in this context (Sedlmair et al., 2014)

Despite the benefits demonstrated and additional information gained through the visual representations in this scenario, one questions the scalability of the current geodemographic process to the multitude of complexities highlighted throughout this research. The fact that the four stage process is complex, time-intensive and sensitive to changes in scale, questions the ability to produce and maintain static classifications when so many options and possibilities for multiple classifications over scale (spatial, temporal

and attribute based SR and SE) and geography will be available. Considering the potential to visualise multiple variables across geography and scale using the frameworks it is suggested that specifically designed visualisation of this type is worth exploring as an alternative to producing and continuing to refine (smart) energy-based geodemographic classifications. This approach termed *noClassification* for the purpose of this research is discussed further in Section 8.2 in relation to the Smart Home Analytics scenario.

Whilst there is more research to be done, this chapter reveals that geography and scale should be considered when variables are selected for geodemographic classification. Continued development of a visualisation tool to aid the geodemographic classification process is seen as important to geodemographic research. It is seen as beneficial to understanding the complexity of the variables and may help to open up the complex process of generating geodemographic classifications to wider audiences. For improved usability a number of extensions to the prototype outlined in the following section (Section 7.7) are seen as necessary to future work.

7.7 Possible Extensions for the Prototype

Throughout the last few sections many possible improvements to the instantiation of the framework in the context of generating a domain-specific geodemographic have been mentioned. These are summarised and discussed in this section. Many of these extensions were classed as ‘should-’ or ‘could-haves’ at the requirements stage, as they were either seen as not essential for the representation of the framework or too complex for development. These extensions reference user stories in Table 4.1 and are referred to in later text.

Hierarchy: A decision was made early in the development process to show all variables at the global level, reflecting the current correlation matrix used in geodemographic examples (as shown in Fig. 2.5) and then allow the user to drill-down to the local level to investigate geographical variation or investigate the effects of scale. This decision results in variables always being the first dimension and secondly there is a choice of geography or scale. A more advanced design could allow the user to have the choice as to whether to divide the data first by scale, geography or variable and each option will display the data in different ways (depending on the second or third dimension). This allows the user to explore the data in multiple dimensions. Being able to switch the hierarchy ordering of levels is implemented in HiDE/HiVE software (Slingsby et al., 2009) as shown in the three images in Fig. 4.2, which highlight different aspects of the same dataset when switching the first and second level of data represented from geography to MOSAIC Group. Potential designs for a three level hierarchical representation are discussed in Section 8.1.

Locality with Scale: A limitation of the current design is that scale (SR) is not compared together with the local statistics as the option allows only scale or geography. The combination of the two would allow investigation of the local and global values across scale and further aid the interpretation and understanding of the effect of aggregation.

Filtering (US#28): Another limitation of the current design is the limited filtering options. Variables can be reduced and increased using the -/+ buttons as well as reordered but additional filtering options were not implemented. Filtering variables by domain, filtering by chosen variables or contracting the matrix when variables are masked (like in the ‘show extreme values’ option) could easily be added to improve usability.

Highlighting/Flagging (US#9): A useful extension would include the ability to automatically detect (by highlighting or filtering to) problematic variables. This could include options or sliders for the user to set the restrictions or thresholds. This could be derived from the global statistics shown in P1. This is a more sophisticated indicator than the current ‘show extreme values’ option. Local geographical variation could also be included, to clearly highlight which variables contain local geographical variation in comparison to global statistics. Variables which are sensitive to scale could be highlighted (both automatically and manually) in order to ensure that the appropriate scale is utilised for analysis or to distinguish variables which relate to phenomena at different scales. This could have overcome the OAC 2001 London ‘Multi-Cultural’ Super Group problem to an extent and could allow for the investigation of classifications with the inclusion of regional phenomena using data at different scales, such as the concentric circle method used by Experian (as described in Section 2.3.3.4). Along with the ability to flag variables to be included or excluded, it would be useful to see whether their relationship with proxy variable(s) is scale dependent or geographically variable and label these combinations as having some scale or geography. In the context of reducing variables, these labels could then assist the removal of variables which are globally close because of local variation.

Support for Analytical Sensemaking: Along with flagging and highlighting options, there is a need to annotate the variables for future reference and the ability to save and reload the state of the application (as addressed in HiDE software (Slingsby et al., 2009) used in Section 4.1). Saving the history in order to support actions such as undo, replay and progressive refinement (Shneiderman, 1996) of the analytical discovery process is important to not only support progressive visual analytics (Thomas and Cook, 2006) but also provenance (Walker et al., 2013). While saving the screen as a static image was implemented in the prototype, with the information needed to re-create the image stored in the file name, an automatic process was not developed. These are fundamental options which would enable a more sophisticated application to be used and tested by others and for different case studies. Whilst the current design is only designed

to be a demonstrator of the framework, a more sophisticated visual analytics system, purpose-built with these options may help to understand more about how the proposals work in action and for that it needs the support for provenance.

Adding New Data: The ability to add new data (US#12) to the tool is necessary to allow it to be used for other purposes. The smart home data (US#13) would be useful to allow detailed profiles to be created. The use of this data introduces time to the application and therefore the extension of the application to temporal scale and attribute scale, where variables can be investigated in a similar way to the spatial scale analysis, but through aggregations and filtering of periods of time or attributes of the variables. The continued investigation of spatial scale is also needed with the analysis of spatial extent, where parallel matrices could be compared in juxtaposition to show differences across two or more chosen SE. Example designs of all these options are discussed in Section 8.1 and Appendix B.10.

Map Marker: A map marker tool could allow areas to be selected geographically for specific investigation, which could automatically trigger relevant or the most discriminating variables for consideration. Such a map marker tool could support the decision process by allowing the variables to be chosen through the map and by recording these selections on the map. Specific geographical areas, for instance the isles of Scilly (see Section 7.3.1), could also easily be investigated separately or even removed from the analysis.

Map Design: In terms of geographical representation, the location-aggregated raster maps are useful but not 100% representative of the underlying data. A more effective approach to the problem would be to aggregate by pre-defined geography using a hierarchy of geographical areas. The geographical space could also be used more efficiently with use of treemaps or cartograms as discussed in Section 2.4.3. The addition of difference maps (and other views showing the difference calculation) would also greatly improve the usability and interpretation of the data relationship when changing the aggregation or calculation of locality.

Statistical Methods/Clustering: Other useful extensions to the current design include the ability to investigate further transformation options (US#21). The effect of transformation on some variables and not others would be useful to represent as this may help to reduce strongly correlated variables (as noted in Section 7.3.3). Additionally, the ability to implement thresholds (US#24) or the introduction of weighting are useful alternatives to transformation. Being able to vary the effect of weighting (US#36), thresholds or transformation interactively and see the effect in real-time is important for understanding these processes. These options all link to the other three stages of the geodemographic classification process, where each stage of the

process is visualised and the decision to choose or remove a variable can be seen in the clustering process and results (US#37-47).

In summary, there are many possible extensions to the prototype with respect to creating a usable and useful variable selection for geodemographics application. Some are more complex than others. These extensions and usability issues have all been identified through the implementation and subsequent validation of the prototype for this scenario, yet the instantiation remains a useful demonstrator of the model as it allows for the systematic representation of the complex parameter space and opens up the design space for continued research.

7.8 Chapter Summary

In this chapter the instantiation and the model are validated through an in-depth investigation of variable selection for the generation of energy-based geodemographic classifications (Scenario 1). The visual analysis of the energy variables and the parameter sensitivity identify that not only is the parameter space complex, but that the multivariate geography of the UK with regard to energy use is also complicated. Visualisation designs based upon the framework are beneficial to understanding some of the complexity. The exploration of the scenario in this chapter helps to address all four research questions outlined in Section 1.2. Many extensions are also highlighted and outlined as suggestions for further development in Section 7.7. Although the prototype is shown to be beneficial in this context, the exploration of the scenario suggests that there is also an opportunity for a noClassification approach offered through interactive and dynamic geographical (g)vPSA. Such gvPSA tools, designed to help analyse geographical patterns in multivariate data, may prove to be more efficient and effective than energy-based geodemographic classifications to profile consumers in practice. This is investigated in the following chapter, which discusses the potential for the framework and prototype designs for visualising variables in two further scenarios of Smart Home Analytics and Survey Response Modelling.

8

Further Applicability of the Framework

This chapter describes further research on visualisation designs and scenarios applicable to the framework outlined in Chapter 6. Section 8.1 explores possible visualisation designs such as adaptations to the mosaic scale view, which would allow the prototype to be extended to different data sets, such as those with temporal or attribute based scale. Section 8.2 describes the noClassification approach to geodemographics in reference to the research in previous chapters and the four stage classification process outlined in Section 2.3.4. Section 8.3 then describes two scenarios where the framework, the prototype design, extensions (Section 7.7) and these potential design ideas are shown to be applicable. The research in this chapter goes beyond the original remit of energy-based geodemographics to show the potential impact of the artifacts generated in this research in other gvPSA contexts.

8.1 Designs for Future Work

As described in Section 5.2 there are multiple dimensions of scale – spatial, temporal and attribute-based SR and SE –, and data transformations can be visualised using such spatial and statistical representation as described in Section 6.2 and Fig. 6.3. The data dimensions discussed in this research – multiple variables (V), geographical variation through local statistics (L), and spatial-, attribute- or temporal-scale (SR and SE) – could be visually represented in many different ways.

In the prototype design, the GlobalMany view for V was used as the default view for variable selection, as a static correlation matrix had already been used in Geodemographic research (see Fig. 2.5). In this section an adaptable three level (L1, L2 and L3) hierarchy is explored, where the level represents the division of the data, whether that is first by variable, geography or scale, e.g. allowing the user to start from different areas of the framework such as from MacroMulti or MicroBi. The hierarchical structure was not used for the prototype design due to tailoring the prototype to the geodemographic variable selection user stories, where variable detail as well as overview of all variables is important. The design of the prototype was therefore restricted to using variable as the default division of the demonstrative dataset. The multi-level hierarchical structure to switch the data division from V to L to SR or SE discussed in this section, is influenced by hierarchical representation research (Slingsby et al., 2010a) and the HiVE/HiDE software (Slingsby et al., 2009), which was used to produce Fig. 4.2 and 4.3.

Three levels are investigated as a maximum in this research as screen space limits possibilities in most desktop analysis situations. This also reflects the three dimensions of Variable, Geography or Scale, which are all seen as equally important to the analysis. The framework described in this thesis (Chapter 6) provides guidelines for multivariate geographical analysis, where geography is clearly an important dimension in all instances, together with dimensions of scale, whether spatial, attribute and temporal, as described in the following section for two additional scenarios. Transformation may also be applicable to the analysis, as shown in the Geodemographics variable selection scenario (described in Chapters 4-7).

8.1.1 Brainstorming Hierarchical Designs

A brainstorming exercise took place prior to the prototype design decisions and development. This involved in-depth thinking of certain cells of the framework – GlobalMulti, MacroUni and MacroMulti. These brainstorming explorations are available in Appendix B.10, where a blank A4 sheet of paper (then later transferred to digital) was divided into different sections relating to the dimensions of Variables, Geography, Scale and Transformation. Scale was given the largest section as it covers both SR and SE for each dimension of spatial, attribute and temporal. Geography was split between spatial and statistical views, as it was noted that both are important in the context of variable selection for geodemographics. Transformation has only a small section as in this instance transformation is seen as a boolean option rather than an investigation of multiple transformation types. This was important in order to reduce the number of possible options to visualise and concentrate the investigation on scale and geography. It

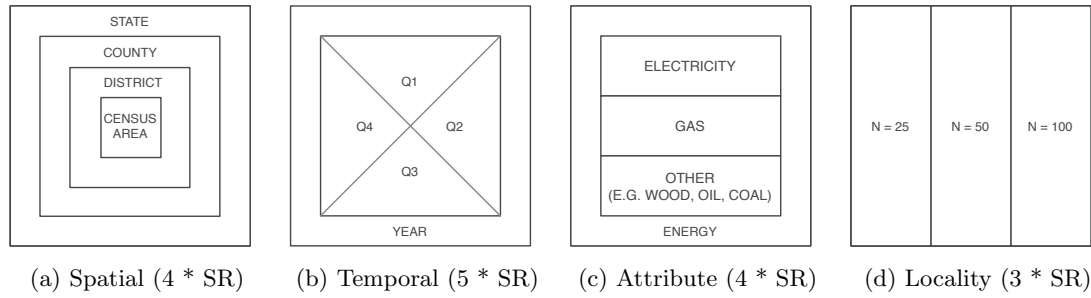


Figure 8.1: Scale mosaic designs. Spatial: statistical boundaries (hierarchical); Temporal: seasons and yearly total (circular / hierarchical); Attribute: types of and total household energy consumption (nominal / hierarchical); and Locality: ordered number of neighbours N (ordinal).

was seen as the least relevant to investigate further as there is already visualisation literature and geodemographic research in this area (see Chapter 2).

The use of the matrix representation was investigated for all these cells with the use of combining juxtaposition and superposition when switching between types of data, e.g. as SR shifts from attribute to temporal to spatial. Superposition is used mainly for the comparison of scale, while juxtaposition is useful for distinguishing variables and representing transformations. Asymmetrical matrices are important for showing geographical and statistical views as well as non-logged or logged representations. Some of the designs particularly those related to SR and Geography were developed for building the prototype to demonstrate and visualise the framework in the context of variable selection.

8.1.2 Multiple Scale Mosaics

The scale mosaic design in particular is found to be applicable to many dimensions of scale. Fig. 8.1 demonstrates the applicability of Scale Mosaics to alternative Scales. Fig. 8.1a shows the hierarchical arrangement of spatial geographical areas, as presented in the prototype design. Representations can be adapted to temporal- (Fig. 8.1b) and attribute- (Fig. 8.1c) based SR or SE through altering the orientation of the splitting of the matrix cell using juxtaposition or containment (Gleicher et al., 2011). Fig. 8.1 shows candidate designs reflecting cyclical, linear and hierarchical arrangements dependant on the data. Increasing the number of variables (V) under comparison allows this design to move from the GlobalBi (Fig. 8.1) to GlobalMulti – as shown in the prototype design in Fig.6.17b. With the scale mosaic design, increasing the number of V reduces the ability to compare multiple scales and a single statistic is presented for the GlobalMany cell of the framework. Candidate statistics are the variance, rank or range of the values associated, e.g. as shown in Fig.6.17d.

A further application of the design is present the neighbourhood (N) parameter from the locality definition in this format, as shown in Fig. 8.1d. This could equally apply

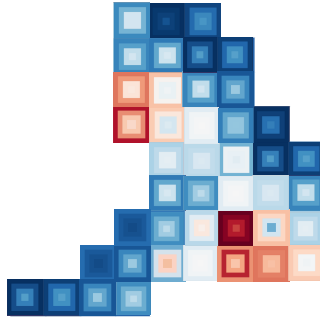


Figure 8.2: Basic design for hierarchical scale mosaic designs within the raster map design of the prototype

to the alternative locality calculations as described in Section 5.1. As these values are associated with the local statistics rather than the global (as is the case in Fig. 8.1a-8.1c). Alternative geographical views are needed in order to present both SR and L, as discussed in the following section. Hierarchical representations can also be used to select geographical area as the first division, rather than by variable. In the brainstorming activity it was noted that if V is at L2 or L3 instead of L1 in the instance of geodemographic variable selection where there are many V, then the number of values in L1 (e.g. geography or scale) need to be limited, for dividing the presentation first by the 9 NUTS1 regions or by the four spatial SR used in the prototype (NUTS2, LAD, LSOA, OA), as discussed in Section 5.4.2.

8.1.3 Locality and Scale Mosaics

Due to limitations on resources, local statistics and variations for N were only calculated for NUTS2 and LAD for the prototype design. A next step to this research is to combine the geographical views (i.e. Macro and Micro levels of L) with the values for scale. Designs for combining geography and scale in the brainstorming activity (Appendix B.10) show the scale mosaic cell designs mapped internally within the geographical view. An example of this is illustrated in Fig. 8.2 where the hierarchical scale mosaic designs and square raster map views used in the prototype are combined. Such map designs could also show cells reporting different locality values, such as in Fig. 8.1d. The combination of locality and scale mosaics is described as a useful extension to the prototype in the previous chapter (Section 7.7 – ‘Locality with Scale’), although further work is needed in order to demonstrate feasibility and usability of this view. Complexity increases for the user especially when comparing more than two variables or where L is large, therefore this view is not expected to be useful for more than a few variables. It is particularly applicable to the UniMacro (e.g. local skewness) and BiMacro (e.g. local correlation) cells of the framework.

8.1.4 Designs for Future Work Summary

This section demonstrates a number of possible design options for future work based on a three-level hierarchical framework, influenced by hierarchical visualisation research at the giCentre (Slingsby et al., 2009, 2010a). Some of these designs are visible in the prototype, but the alternatives demonstrate that the prototype design is adaptable to future work and continued development. These designs, especially the multiple scale mosaic options for temporal or attribute based data, are shown to be applicable for the additional scenarios in Section 8.3.

8.2 Dynamic Geodemographic noClassifications

The research outlined in this thesis has investigated the process of generating domain-specific geodemographic classifications and the importance of data visualisation in this process. Visualisation is shown to be appropriate to this complex and time-intensive geodemographic process, particularly in the variable selection process demonstrated through the evaluation of the prototype in the previous chapter. Yet research also shows that the future of energy-consumer profiling will have more information and be data plentiful due to the adaption of the energy sector to a smarter more efficient system. When proposing energy-geodemographics for future energy analysis, the need and capability to produce energy-based geodemographic classifications is questioned. Whilst the new data sets are anticipated to produce richer and more discriminating energy profiles, the ability to add and deal with frequently changing, near real-time data within the geodemographic classification process is unlikely. The fact that the current process is shown to be complex, time-intensive and sensitive, questions the ability to produce and maintain static classifications when so many options and possibilities for multiple classifications over scale (spatial, temporal and attribute based SR and SE) and geography will be available.

Although irregular static geodemographic classification is expected to still be extremely useful in the context of improving energy consumption understanding and consumer behaviour (as discussed in Chapter 2) a more adaptable dynamic approach is likely more manageable where data is plentiful and frequent updates needed. Geographical (g)vPSA tools designed to help analyse geographical patterns in multivariate data can allow variables relating to energy consumer characteristics to be grouped, filtered and segmented, similar to the goals of geodemographic classification, but allow for new variables to be added, potentially in real-time.

Specifically-designed interactive visualisation tools to investigate the parameter space through gvPSA could allow selected variables to be reduced, grouped and dynamically ‘classified’ using statistical methods such as clustering and PCA. Variable selection is

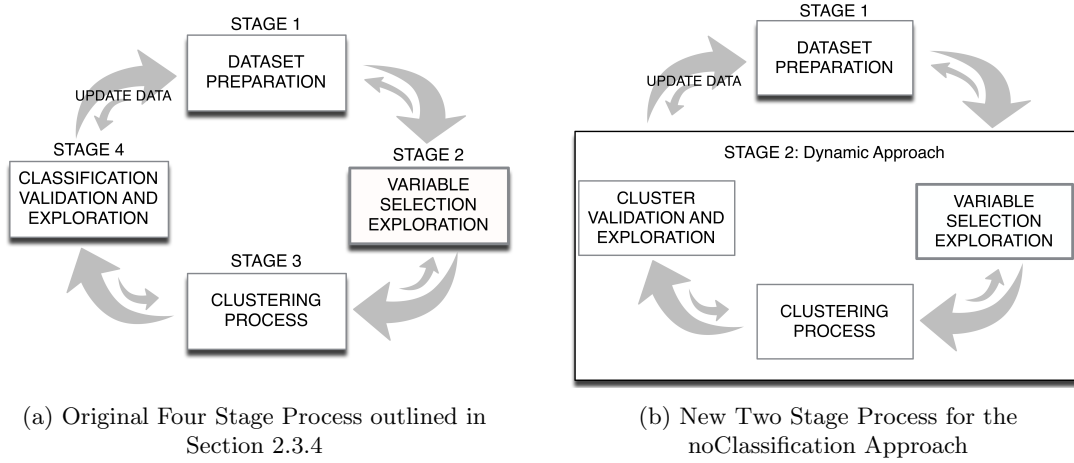


Figure 8.3: Updated four stage process, removing the static classification and embedding the noClassification approach through dynamic visualisation methods and connected processes

still a vital component but this process is combined with Stages 3 and 4 (see Section 2.3.4). The three later stages of the four stage process are replaced with a dynamic visual representation rather than resulting in static clearly named and profiled classification results. This change to the original diagram is illustrated in Fig. 8.3. This approach is adaptable to new data and more aligned with the way in which data is now being produced – such as sensors monitoring energy consumption in real-time.

The dynamic approach outlined here is not seen as a replacement to static domain-specific geodemographics but as an extension to the currently complex, time-intensive process, especially in situations where data is plentiful and classification results need to be flexible and adaptable. This more dynamic approach to ‘classifying’ data sets is discussed further in the following chapter and reflected on in the following section where the use of the framework and designs is applied to two additional scenarios. The framework for multivariate geographical visual analysis described in Chapter 6 and gvPSA itself is seen as important in both the static and dynamic approach.

8.3 Additional Scenarios

Whilst the functionality of the prototype was designed within the context of geodemographic variable selection, the framework is intended to be flexible and adaptable to other problem areas, which would benefit from the comparison of multivariate data across scale and geography. Two scenarios to demonstrate the further applicability of the framework are described in this section. The first refers to the smart home project discussed in Chapter 3. The second is based on a current research project ‘ADDResponse’ investigating survey nonresponse bias.

8.3.1 Scenario 2: Smart Home Analytics

The domain knowledge acquired from the smart home project described in Chapter 3 generates a second scenario for the framework. The outcome of the requirements gathering workshop identified a real need for discovering new knowledge through customer data analysis in order to communicate and engage with customers, build trust and subsequently improve products and services (see the five outcome themes from the workshop in Section 3.2). Sophisticated analysis of smart home data is seen as critical to understanding energy consumption habits and consumer behaviour.

8.3.1.1 Smart Home Analytics and the Framework

The need to know the “*who, what, when, where and why of energy consumption*” demonstrates that many variables are needed for the smart home analytics scenario. This scenario refers to appliance-based consumption data over time (as investigated in the four smart home prototypes in Chapter 3), linked by geographical location to socio-demographic variables about consumer characteristics (as discussed in Section 4.3. Although the trial and model data sets available for the smart home project (see Section 3.3) have only limited demographic and geographical variables, it is likely that future smart home data sets (whether real or through improved models) will contain variables related to *who* people are, *what* they use, *when* they use it and *where* they live.

Improved analytical processes and techniques will allow for the better understanding of *why* energy is being consumed. The concept map produced for the smart home project (Appendix B.4) demonstrates some of the potential links for data analysis. While the regular generation of static energy-based geodemographics is debated in this scenario due to the time-intensive process and near-real time data (see Section 8.2), visualisation is still seen as fundamental to understanding the complexities of the variables and for gvPSA. When considering the possible dimensions there are hundreds of possibilities, especially in relation to scale, and the parameter space becomes vast. The use of the framework for comparing multivariate data across scale and geography is therefore seen as beneficial in this scenario.

Firstly in regards to scale there are a number of options to aggregate (SR) and filter (SE) the data based on time, attribute and geography:

Temporal Scale: The smart home data can be aggregated and/or filtered based on many periods of time, e.g. by second, minute, hour, day, week, month, season or year. Certain appliances are more sensitive to temporal aggregation than others; kettles, for example, use a lot of energy in a very short period of time (seconds-minutes), while washing machines use a lot for a longer period of time (minutes-hours) and other appliances are on all the time or fluctuate power use depending on need, such as

appliances which have a standby option. These different types of appliances can be identified in the model created for the smart home project, where nine classes (A-I as demonstrated in Consumption Signatures see Section 3.5.2.2) of appliance are defined based on power demand and duration of use (Gruber and Prodanovic, 2012). When data is aggregated the differences between appliances can be lost and therefore it is important to be able to explore the data across multiple temporal scales. The ability to show how data changes with temporal aggregation can also help the energy company define at what scale to store the raw data¹ and how long for, but also at what scale the data should be represented to different users, for example business stakeholders, analysts or customers.

Attribute Scale: There are many options for aggregating and filtering the data based on attribute, such as by fuel type, appliance type and household type. The ability to segment the data in multiple ways is essential for the analyst to be able to gain insights and improved knowledge of energy consumption patterns in order to better manage supply and demand, and optimise the customer experience through improved products and services. Such services could involve creating tailored tariffs or new products which move smart homes beyond energy towards comfort and security.

Spatial Scale: The variation of the geographical differences between and within energy variables is explored in this thesis (particularly in Section 7.2) and the results demonstrate that there is much complexity involved, even for the relatively few variables investigated. With many more data dimensions possible, being able to group and filter the available variables by geographical scale will allow the geographical patterns of energy consumption to be explored and better understood. In terms of SR, the household level data can be grouped by street, neighbourhood, borough or region and multiple geographical scales can be visually compared, as demonstrated in the prototype, through the use of the scale mosaic design. These aggregations can help to discover phenomena, which occur at different geographical scales. Aggregation of the data also provides data privacy and therefore can be used for presenting to the industry or for improved customer feedback, such as allowing users to be compared to others in their neighbourhood. Additional analytical comparison tasks such as “*compare City A to City B*” (an idea from the requirements workshop described in Chapter 3) are made possible through filtering SE.

Locality: The addition of locality allows for the investigation of local geographical patterns within and between variables. The framework allows for many variables to be displayed on one screen (GlobalMany) with the incorporation of geography (Macro/Micro) and the ability to investigate the geographical variation in more detail when comparing multiple variables (MacroMulti and MicroMulti). For detailed investigation of variables

¹Presuming permission to store and analysis the data has been acquired

there is also the Bi- and Uni- options. As there are so many dimensions to consider, the use of the hierarchical levels, as discussed in Section 7.7, will be useful for visual analysis, in order to switch the dimension of interest from variable to geography or scale (attribute, temporal and spatial SR and SE). This will enable patterns of energy consumption to be discovered based not only on geographical location and scale (as demonstrated by the prototype), but through variations of time or attribute; addressing the aspiration: *“I would like to know how electricity and gas use changes with age (lifestyle, life cycle, group, segment)”* (from the requirements workshop described in Chapter 3).

The use of grouping and filtering as well as reordering are seen to be beneficial, as demonstrated by the framework prototype and all 4 smart home prototypes (Chapter 3). The fact that ‘Smart Home HeatLines’ and ‘Consumption Signatures’ were so well received by the analysts was partly due to the fact the application allowed a lot of data to be displayed in one view. The magnitude and complexity of the data is masked through aggregation and filtering by time and attribute. The visualisation possibilities for smart home analytics through the use of the framework could produce a knowledge building tool similar to that of ‘Consumption Signatures’, which prompted the response: *“I could imagine ... just taking a week off and just letting your curiosity dive in and out.”* The addition of geography and demographics allows many more patterns to be discovered, with the ability to *“slice and dice”* the data, to discover *“typical patterns/predictions”* and *“gain nuggets of knowledge”* about *“why people use energy”*. Such a visualisation discovery and gvPSA tool, potentially with the inclusion of optimisation options (such as those in the smart home model) and mechanisms to perform *data sculpting*, can also allow energy companies to better understand supply and demand and start to *“support the grid”* through understanding the consumers.

8.3.1.2 Smart Home Analytics Scenario Summary

The smart home data is rich with variables and will increase in quantity and resolution. This provides many possibilities for improved data analysis. The framework is shown to be applicable to the scenario through the fact that many variables need to be compared across multiple dimensions of scale (temporal, attribute and spatial) as well as over geography. The framework can provide guidelines for the complex parameter space to be visualised. The creation of a visual analytics tool for knowledge discovery through gvPSA could help to unfold some of the complexities associated with energy consumption in this context, including variations in population characteristics as well as geography and time. This will provide a mechanism for the continued investigation of energy consumption through better understanding of the smart home. Improved analysis through visualisation can lead

to new knowledge on energy consumer behaviour and habits, and enable better tailored services and products for consumers.

8.3.2 Scenario 3: Survey Response Modelling

Efforts to use auxiliary data to explore, explain and predict nonresponse bias in social surveys are being funded in the UK. The Economic and Social Research Council ‘ADDResponse’ project (to Research, 2014) is using multivariate data at multiple scales to understand the likely response levels in different areas, in order for these to be accommodated in survey design. The objective is to use multivariate data to develop more representative and less biased survey responses. The project is a collaboration between the Centre for Comparative Social Surveys and the giCentre at City University London, together with the Department of Statistics at the London School of Economics².

8.3.2.1 Survey Response Modelling Overview

Nonresponse to general population surveys is a considerable problem with most response rates only reaching 50% or less (Massey and Tourangeau, 2013). The research project aims to use auxiliary data to identify geographic and demographic patterns in UK survey responses in order to improve future survey data through better weighting and enable the generation of more robust data to better understand public attitudes and behaviour. The outcome is expected to greatly enhance general social surveys results in the UK and be applicable internationally (ADDResponse, 2014). As it is recognised that neighbourhood contextual data can provide relevant information about response rate (Campanelli et al., 1997), the project is focusing on three strands of work³.

Strand 1 involves scoping available auxiliary data sources to append to the UK European Social Survey (ESS) sample data. This includes three types of data sets:

1. Small area administrative socio-demographic neighbourhood statistic data sets, e.g. the Census and DECC variables used in this thesis at OA and LSOA levels
2. Commercial household level consumer marketing data. e.g. from Experian⁴ and CallCredit⁵
3. Geocoded information from the Ordnance Survey (OS) on physical locations to allow all the data sets to be combined and analysed geographically

Strand 2 involves identifying which auxiliary variables correlate with nonresponse bias in the sample and answering the following research questions (ADDResponse, 2014):

²For more information on the project and principal contributors see the project blog at: <https://blogs.city.ac.uk/addresponse/>

³Explained in more detail on the project blog and case for support document available at: <https://blogs.city.ac.uk/addresponse/>

⁴Experian Marketing Services: <http://www.experian.co.uk/marketing-services/>

⁵<http://www.callcredit.co.uk/>

- To what extent can auxiliary data be used to predict sampled households' propensity to respond?
- How are any auxiliary variables identified as predictors of survey response associated with different survey outcomes?

For this investigation of variable relationship it is proposed that analysis will focus on the correlation between auxiliary variables and themes (such as wellbeing and fear of crime) in the ESS core questionnaire. In order to be a good predictor, the variable needs to be tested for its usefulness when varying the level of aggregation of the data or variability of the geography (ADDResponse, 2014). Geographically weighted regression (GWR) will be used to investigate geographical comparison at different scales of resolution (Fotheringham et al., 2002) and visualisation is proposed to explore, present and communicate the geographically varying results (e.g. Slingsby et al., 2011) and relate them to the most significant local explanatory variable.

Finally, Strand 3 will use the chosen auxiliary variables to develop improved weighting for nonresponse bias and test these on the ESS sample.

8.3.2.2 Survey Response Modelling and the Framework

In this scenario, outlined in the previous section, there is a need to successfully combine and compare multiple measurements over multiple scales with different levels of aggregation. The first two strands of work outlined in the previous section are similar to the first two stages of the four stage process for generating geodemographics (see Section 2.3.4); where Stage 1 evaluates relevant data sources for feasibility and quality and Stage 2 combines these for variable exploration, where relationships and correlation are investigated and differences to variable scale and geographical variation are evaluated. The parameter space for visual comparison necessary for this scenario is well suited to the framework as it must take into account many data sets containing many variables (i.e. $V = \text{Many}$), scale (SR and potentially SE) and local geographical variation (L). While the prototype discussed in the previous chapter would need to be redesigned for this type of analysis, there is great potential for this project to further demonstrate the utility of the framework.

To complement the use of GWR suggested above, the use of locality enables local summary statistics (such as local correlation) to be calculated and used in the visual representation and comparison. A global overview of many variables (GlobalMany) allows variables with correlations to survey responses to be detected visually. Given the number of variables, interactive reordering of the matrix (or other form of visualisation deemed beneficial) will be needed for pattern detection. Grouping and filtering the variables will also be critical for data reduction and to allow as much as possible to be shown on one screen.

The inclusion of SR and SE based on attribute and time as well as spatial scale is necessary for the analysis of the data in this scenario. The use of the hierarchical levels (see Section 8.1) would be particularly useful for this scenario; for example, investigation by domain, attribute type, data source, nonresponse type, geography, time, or by similarity of variable profile (influenced by the SmartHome HeatLines order by profile option, see Section 3.5.2.1). For geography, an efficient map design for level 1 of the hierarchy and a ‘map marker’ tool (see Section 7.7) would allow geographical patterns in the data to be discovered more efficiently than in the current prototype. With the use of aggregation and filtering at the standardised stage (StR and StE, see Fig. 5.3) the differing visual design options for scale mosaics at geography, attribute and temporal scales (discussed in Section 6.2) are relevant and may prove useful. For this scenario the extensions mentioned in Section 7.7 relating to automatically detecting ‘interesting’ variables based on a set of criteria, highlighting and flagging variables, making annotations for future reference as well as saving and loading states would all be necessary in order to ensure that the variables chosen could be rediscovered, justified and presented.

While aggregating and filtering the variables may be enough to enable visual comparison and insight, it is likely, due to the sheer number of variables that are expected, that some form of statistical sampling or dimension reduction will also be needed; for example PCA or to include geographical variation there is also geographically weighted PCA (Harris et al., 2011). Clustering may also be considered to identify auxiliary variables relating to survey responses. The ability to automatically detect patterns and allow the user to have visual confirmation could prove useful in this scenario. If the ‘perfect’ variables to identify response rates are discovered through this research project, then profiling the population based on response rates to surveys, using geodemographic classification methods, could be useful to research in the area. A static classification, grouping the characteristics of the population based on survey response rates, would be beneficial in this case as it is unlikely the classification will change often and therefore would not be subject to the issues of frequently keeping the variables up-to-date, as in the previous scenario. As the statistical methods used in geodemographic classification can create uncertainties and lead to misinterpretations, there is a need for visualisation to enable these statistical transformations to be more transparent with fluid transitions and links back to the raw data.

To confirm that the chosen variables perform well across scale, the scale mosaics design for GlobalMulti, as shown in the prototype, could prove beneficial. In order to investigate the variance of variable scale across local geography, the continuation of visual designs for combining both scale and locality (i.e. when $L = \text{Macro}$ and Micro) is needed, as discussed in the prototype extensions (Section 7.7). The fact that this scenario compares so many

more variables than in the prototype means that even at the GlobalMany stage the data will need to be aggregated or filtered in order to be able to visualise everything on one screen. This indicates that the options which are discussed as visual possibilities options for the framework in MacroMany and MicroMulti (less data items, high data density, more pixels – see Table 6.3), are likely to also be needed for GlobalMany. The framework may need continued research for such a scenario. This is not seen as a failure but as an indicator that there is research to follow and further use cases need implementing to test the model: *“It may be instantiated out of intuition and experience and only as it is studied are we able to formalize the constructs, model and methods”* (March and Smith, 1995, pp.258).

8.3.2.3 Survey Response Modelling Summary

The ‘AddResponse’ project discussed in this section involves combining many variables over multiple scales with the consideration of geographical patterns in order to detect patterns in survey nonresponse rates. With better understanding of the variables that affect nonresponse bias, a more robust weighting can be deployed to predict and explain survey response rates. As project combines multivariate analysis with multiple dimensions of scale and consideration of geographical variation it fits with the framework described in Chapter 6. Although the instantiation of the framework does not meet the exact requirements needed in this scenario, a number of functions, features and extensions (Section 7.7) and the multiple scale mosaic designs (Section 8.1.2) are identified as useful. In particular, the use of aggregation and filtering in combination with hierarchical levels can be used to efficiently change the visual representation and quickly gain visual insights. The explanation of this scenario in this section demonstrates the utility for the framework outside of the energy domain. While there are still many visual possibilities for presenting this scenario to the user, the framework can act as a guideline for these decisions.

8.4 Academic Feedback

The framework and prototype have been presented (mainly through videos and stills) at a number of conferences and invited talks (see pages xvi and xvii) where feedback has been very enthusiastic. When presented to design students at the University of Applied Sciences in Potsdam, for example, the prototype was seen as a really positive step to help understand the complexities of multivariate data analysis, as often visualisation designers try to reduce data to the simplest of measures in order to achieve mass understanding of the statistics and the complexities or sensitivities of analysis are not expressed. The inclusion of geography in the multivariate comparison was seen as unique and beneficial for improving the understanding of how multivariate data can vary geographically. The

prototype was seen as an educational tool, which with continued development, could have potential for demonstrating statistical methods through visual representation in teaching. This, in combination with feedback from Chris Gale (see Section 7.5) and a discussion with Michael Sedlmair to confirm the overlap with vPSA (Sedlmair, 2014), encourages continued work in the area in order to improve people’s understanding of the complex parameter space and educate users on the possibilities of gvPSA.

8.5 Chapter Summary

Further work applicable to the framework in addition to the scenario of variable selection for geodemographics is outlined in this chapter. This includes potential designs for future work as well as the potential for a more dynamic non-static geodemographic classification process utilising the benefits of visualisation, termed noClassification. The research continues to demonstrate that there is not just a need for the framework within variable selection for geodemographics but that the framework is relevant to other scenarios. The smart home scenario shows future need, whilst the ‘Survey Response Modelling’ scenario shows current research in which the framework can be utilised. The descriptions of scenario 2 and 3 demonstrate the broader applicability of the framework and present a need to not only compare spatial scale – as focused on for the prototype – but that attribute and temporal scales are equally important to multivariate geographical analysis. The framework, in combination with some of the new designs (Section 8.1), prototype designs (Chapter 6) and possible extensions (Section 7.7), is shown to be appropriate for the data and analytical tasks in each scenario.

All three scenarios (including Scenario 1) demonstrate the need for the visual analysis of the geographical variation of multiple variables in combination with scale. While there are areas in which further research is needed, the framework is shown to be useful and applicable in context. It is evident that there is a need to continue the research addressed in this thesis in order to help support and further explore the potential for gvPSA and variable selection, where geographical variance is important, in both current and future practices. Through the investigation of the utility of the framework (in this chapter and the previous) it is evident that the creation of a visual analytics tool for knowledge discovery through gvPSA could help unfold some of the complexities associated with multivariate analysis across scale and geography. In the context of the scenarios, this can enable insights and new knowledge about patterns in energy consumption, consumer habits and behaviour – through areal data sets (Scenario 1) or smart home data (Scenario 2) – or assist in the discovery of variables related to nonresponse in surveys (Scenario 3). The following chapter continues the discussion and links the framework, the prototype and these scenarios to the research context and the research questions outlined in Chapter 1.

9

Discussion and Conclusions

This final chapter discusses the research outlined in Chapters 3 - 8 with reference to the research aims and objectives, the motivational and research questions outlined in Chapter 1 as well as the literature identified in Chapter 2. The research scope, contributions, limitations and options for future work are also summarised.

9.1 Summary of Thesis

The research in this thesis covers the broad topic of visualising UK household energy consumption characteristics. The research began with the broad investigation of the future of (smart) energy data analysis and visualisation requirements with energy experts, using creativity techniques to encourage creative thinking (Chapter 3). This resulted in the development of four smart home visualisation prototypes, which were evaluated for creativity of design and creativity of the design process. The industry's need to be able to profile energy consumers by *who* people are and *where* they live initiated the investigation of generating a geodemographic classification in the context of energy. The investigation of the process (Section 2.3 and Section 4.2) lead to specifically focusing on the visual comparison of candidate variables with the consideration of data scale and geographical variation. The parameter space in this context was shown to be particularly complex and lead to the further proposal of a new theoretical framework to enable the visual representation of multivariate data with the inclusion of geography and

scale (Chapter 6). The feasibility and effectiveness of the new framework was demonstrated through the building of an interactive visualisation prototype, in light of the needs of the geodemographic variable selection process (Section 6.3).

The utility of the framework was validated through the in-depth investigation of geodemographic variable selection (Scenario 1). The visual exploration of the variable selection scenario (Chapter 7) provides evidence that sensitivities occur when introducing variations in geography and scale into the comparison. The exploration for the scenario demonstrates that visualisation can enable the increased understanding of the key elements of geography and scale when selecting variables for domain-specific geodemographic profiling. The research exposes the complexities within multivariate geographical data and shows that visualisation can help to improve the understanding of household energy consumption patterns. The second Smart Home Analytics and third Survey Response Modelling scenarios along with further designs (Chapter 8) demonstrate the broader applicability of the framework and the wider relevance of the complexities of scale and geography in multivariate geographical analysis.

The approach in this thesis draws on the terminology and methods of evaluation from design science research (DSR). The research results in a series of contributions in the form of DSR artifacts. A new *model* represented by the framework. An *instantiation* of the model in the form of an interactive prototype, with new *methods* of visual designs (scale mosaics) and techniques (creativity) for analysis. The utility of the framework and its instantiation also demonstrate the potential for the new *constructs* of gvPSA and noClassification in context. This research opens up a new design space and outlines possibilities for using visualisation in new ways.

9.2 Scope of Research

Whilst the research drawn upon for this thesis covers many academic areas, the scope of the research is restricted to the creation of the theoretical framework and validation through the building of the prototype (Chapter 6) with reference to visualisation design criteria and the demonstration of its utility through the exploration of variable selection for geodemographics (Chapter 7) and the potential applicability to two additional scenarios in theory (Chapter 8). The research is grounded in the needs of the energy industry and the context of visualising energy consumer characteristics.

9.3 Research Discussion

9.3.1 Creative Domain Exploration

Actively encouraging creativity in the Requirements Workshop for the smart home project (Section 3.2) meant that the discussions were not fixed on the needs and possibilities of

the new smart home data, but lead to wide open discussions about the industry and a vast array of aspirations for data analysis and visualisation possibilities. This rapidly increased the designers’ understanding of the problem domain as the activities helped to “*push domain experts to discuss problems, not solutions*” (Sedlmair et al., 2012, pp.2436). Many of the creative ideas were classed as unsuitable or unfeasible for the smart home project itself and the additional time taken to sort through the ideas could be classed as wasted. Yet it is evident from the research reported in this thesis, that the workshop led not only to four smart home prototypes being built related to the available smart home datasets (with three of the four appropriate for the analyst’s needs), but the knowledge and ideas generated during the workshop greatly influenced further ‘channels’ of research (Wood et al., 2014).

The explicit use of creativity techniques not only resulted in a vast array of ideas, but helped the project team work together, communicate, share experiences and establish trust. Reflection on the use of these techniques led to the IEEE TVCG publication (Goodwin et al., 2013) and begins the investigation of introducing aspects of the creative design process into visualisation design, which is currently dominated by engineering design processes (Vande Moere and Purchase, 2011; McKenna et al., 2014). Some creative design elements along with aspects of design science are drawn upon for this research. Rather than focusing on the use of creativity techniques in visualisation design, the research focuses on the aspirations identified in the workshop and exploring alternative visualisations for energy consumption analysis. Many possibilities for the future of energy analysis, identified during the initial workshop, link to the first two motivational questions which began the research: “*MQ1: What is the future for household energy analysis?*” and “*MQ2: What value can be derived from energy consumption data through data analysis and visualisation?*”

The Requirements Workshop exposed a clear need for profiling consumers based on typical traits, in order to better understand energy behaviours. The first stage of the ‘Plan of Action’ for smart home analysis “*Discover: find out where energy is used*” emphasises that better understanding of consumer characteristics are needed in order to reach the second and third stages of analysis “*Displace Consumption*” and “*Reduce Energy Production.*” The need for reliable and trustworthy analysis is highlighted through a quote from one of the analysts: “*The better Stage 1 is, the better Stage 2 and 3 will be.*” The industry need for profiling consumers based on typical traits, leads to the final motivational question (outlined in Section 1.2): *MQ3: Is there a need for an energy-based geodemographic classification?* The need for an energy-based geodemographic classification is demonstrated in the fact that the general geodemographic groups tested do not segment the population well enough based on

electricity use (Section 4.1), that energy consumption is known in academic literature to be *“highly geographically and socio-economically disaggregated”* (Druckman and Jackson, 2008, pp.3167) as well as the evident need for the better understanding of energy consumers based on who they are and where they live (Section 3.2).

9.3.2 Energy Profiling: static and dynamic approaches

Reducing energy consumption is a goal for the government, industry and many residents. Profiling consumers based on typical traits can assist in the understanding of customers and therefore allow services, tariffs, products and even technology to be better targeted. The specific use of geodemographics for profiling is suggested in this thesis because of the need to understand geographical variation in habits and behaviour. Energy use is shown to vary geographically (Druckman and Jackson, 2008) and including this variation in the data analysis will enable services to be targeted by location. The use of neighbourhood or community comparison and social norms are shown to be beneficial to energy consumption reduction (Allcott, 2011), yet the types of services offered to urban residents may be different from those selected for more rural or dispersed residents. Community based schemes in urban locations, such as Tidy Street (Boucher et al., 2012), can be more intensive and engaging due to close neighbourhood proximity, compared to those which incorporate a wider neighbourhood comparison, such as the postal campaigns run by OPower (Allcott, 2011). The comparison of which variables affect energy use and how these variables differ geographically is therefore particularly important for profiling energy use nationwide.

The investigation of the process of generating a geodemographic classification and implementing the prototype demonstrates that geodemographics are not only complex to generate, but are sensitive to the variables chosen as well as their scale and geography. Adding variations of scale as well as geography to the variable selection process creates a complex parameter space for analysis (PSA). Visualisation tools and techniques to make sense of the sensitivity of the selection are seen as key to understanding the complexity. As visualisation and parameter sensitivity analysis are both shown to be key factors in vPSA (Sedlmair et al., 2014), this research opens up the parameter space to geographical variation through geo-visual analysis, i.e. gvPSA. The prototype built to demonstrate the framework supports gvPSA and can guide us towards a more effective consideration of geography in multivariate comparison, which can help to explain, or least help to understand, energy consumption characteristics.

The research in this thesis demonstrates the need for a more visual approach to the geodemographic classification process (discussed in Section 4.2) compared to the relatively static approach in which classifiers are currently built. This compliments a

trend in geodemographics towards more local and domain-specific classifications (Singleton and Longley, 2009b) allowing the process and its complexities and sensitivities to be better understood by a wider audience. Although the research demonstrates that visualisation is beneficial to the geodemographic process, the need and capability of producing smarter energy-based geodemographic classifications in the future is questioned. The geodemographic classification process is time-intensive and even with the inclusion of visualisation it must still involve a human in the process. Adding and frequently updating new variables as well as taking into account phenomena occurring at different geographical scales may result in many more classifications being created, and keeping these up-to-date and comparable will become extremely difficult. Further research (Chapter 8) presents the alternative noClassification approach through which well-designed visualisation is suggested to segment customers into typical traits, like in geodemographics, but to allow for new variables to be added – potentially in real-time – as well as allow for scale and geographical variability to be visualised. This argument is supported by ‘OACExplorer’ which was built to visualise the variables and explore the uncertainties in geodemographics (Slingsby et al., 2011) and the research with the energy analysts (described in Chapter 3), which demonstrates the opportunity for real time engagement with smart home variables through interactive visualisation.

Specifically designed gvPSA tools to allow the more rapid and comprehensive consideration of the issues of scale and geography, which are addressed through this research, can therefore help to support both analysts making geodemographic profiles – and there will be more of these as data increases and it becomes easier to make them – as well as analysts considering data interactively as an alternative to profiling, e.g. the noClassification approach as discussed in Section 8.2 and in the Smart Home Analytics scenario (Section 8.3.1). It is suggested through this research that specifically designed visualisation to aid gvPSA is useful for the future of energy consumption analysis and that the noClassification option could prove to be a beneficial alternative to domain-specific (static) geodemographic classification in this context.

9.3.3 Changing Channel from Variables to Process

The first research question (RQ1) – *Which demographic or socio-economic variables should be combined with energy consumption variables to enable characteristics of UK household energy market to be identified?* – was initially investigated through the literature (Chapter 2), continued during exploratory analysis (Section 4.1) and through the investigation of possible candidate data sources relating to energy use (Section 4.3). The data variables available are constantly changing with new and potentially very

useful variables becoming available during this research, such as the NEED dataset¹. During the research it became evident that the variables currently openly available for analysis are only a fraction of those available within the industry and more importantly, with improving technology, data is growing in abundance and many new variables, offering more detailed analysis than ever before, will soon be available. With the changing landscape of the industry in mind, a list of currently available candidate variables for this point-in-time to answer RQ1 is shown to be less important than the improvement and transparency of the variable selection process itself, where visualisation shown to be useful (Section 4.2).

The 78 variables discussed in Section 4.3 represent possible candidate variables to investigate for classification, and have been examined in the validation of the framework in Chapter 7. While some variables are shown to offer little in terms of discrimination or explanation (and some are even suggested for removal) exactly which variables to classify in order to discover patterns in the market remains open for continued discussion. This open answer to RQ1 reflects a shift in the research from investigating the individual variables themselves to the discovery of the need for visual tools and techniques in order to aid the user in understanding the complexities of the variable selection process and to allow the sensitivities of the variables selected to be better understood. The research channel (Wood et al., 2014) moved towards the investigation of the decisions made and techniques used during the variable selection process and the inclusion of data visualisation, as this can provide rich, interpretable answers to such complex problems.

9.3.4 Visualising Geodemographic Variable Selection

The detailed investigation of the process of generating a geodemographic classification has been made possible through openly available geodemographics with published and detailed methodologies of the process (e.g. Harris et al., 2005; Vickers et al., 2005; Gale, 2014b), as well as vast technological improvements and the availability of capable software (such as the R project) to allow users to create bespoke classifications (e.g. Singleton and Longley, 2009b; Goulvent, 2012; Adnan, 2011). Specific tools have also been developed to aid the statistical and iterative cycle of the clustering aspect (Adnan, 2011; Singleton, 2012). Yet the investigation of the process (Section 4.2.1) shows that there are many unanswered questions, and the complexity of the decisions and process needs further research. For the energy industry in particular, methods for creating geographical profiles of energy consumers, which take into account the *who* and *where* and potentially the *what* and *when*

¹The ‘National Energy Efficiency Data-Framework’ (NEED) was released by DECC in March 2014. This contains anonymised household level data including relevant energy-based statistics (floor area, energy efficiency band, type of property, year of build, boiler type, main fuel type and gas and electricity consumption). This is available at: <http://bit.ly/1x9LaK3> and would be useful for future extensions to this research

of energy use, must be adaptable, flexible and more transparent to allow for new datasets to be added as the industry changes from a static grid to a smarter more sophisticated and modernised system.

In answer to the second research question (RQ2) – *How can data visualisation aid the variable selection process?* – the design science approach used in this thesis delivers strong evidence that visualisation can greatly benefit the variable selection process and the transparency of the decision process. The efficient use of visual variables such as colour and position, coordinated interactive views, and quick transitions from an overview down to the variable detail are shown to be beneficial. The investigation of the decisions made, the processes used and the flow of the data from input to output prior to the clustering stage is shown to be extremely complex and difficult for a non-expert (Section 4.2). Confidence in the data variables as well as statistical results are necessary for there to be confidence in the final classification. In the generation of geodemographics there are statistical processes, such as data transformation and clustering, which can benefit from the power of visualisation to aid user understanding. Furthermore there are complexities and sensitivities relating to the choice and scale of variables.

The importance of geographical variance as a specific dimension when choosing geodemographic variables is also emphasised through the increased discussion for “*bringing geography back into geodemographics*” as a topic in the academic literature (Callingham, 2007; Singleton and Longley, 2009b). Although in the past geodemographics have been produced with only static visuals, as shown for OAC 2001 (Vickers et al., 2005) and OAC 2011 (Gale, 2014b), there is evidence that interactive visualisation can greatly benefit the feature selection process in clustering (Seo and Shneiderman, 2002, 2005). The inclusion of geographical variation – such as through the use of locally weighted statistics – is argued in this research as being necessary for meeting the specific requirements of geodemographic variable selection. If visual design decisions and local statistics are used efficiently, the variables which may be classed as ‘troublesome’ for the clustering process – for example variables which are heavily skewed or those which are strongly correlated – could be quickly identified and further investigated, and the details of geographical variation can be used in the decision-making processes. The investigation of the energy variables in the variable selection scenario (Section 7.2) demonstrates that multivariate comparison through visualisation is beneficial in quickly identifying specific variable attributes (i.e. relating to distribution, correlation or geographical variation), which are important to consider during the variable selection process.

9.3.5 Exploring Sensitivities through PSA

The identification of the sensitivities and complexities involved in the process of variable selection leads to the investigation of the third research question (RQ3) – *What are the sensitivities and uncertainties associated with variable scale, geography or transformation in multivariate comparison?* For variable selection, scale is shown to be important and is investigated in detail for the creation of the framework (Chapter 6). Four stages of scale of resolution (SR) and scale of extent (SE) are shown to be relevant to the variable selection process – input, standardise, locality and output – with regard to spatial-, temporal- and attribute-based scale (see Section 5.2 and Fig. 5.3). The outcome of classifying data at different SE has been investigated in the literature, through the creation of LOAC, a fundamentally different classification for London than for OAC, created for the whole of the UK (Petersen et al., 2011). Whilst variables with different SR and SE are known to be used in commercial geodemographics (Harris et al., 2005), little is published on their methodologies. It is known that geodemographic methodologies in the UK are beginning to be adapted (e.g. CACI, 2014) in light of the increasing availability of open data at different scales and the unknown future of the UK Census (Martin, 2006). The investigation of how SR and SE affect geodemographic profiling can therefore prove beneficial to future research in the domain.

The investigation of the sensitivities and complexities of data scale, geography and transformation for RQ3 lead to the creation of the new theoretical framework (Chapter 6) for enabling the visualisation of multivariate data across scale with the inclusion of local geography. The investigation of the sensitivity of scale, geography and transformation was investigated through varying parameters during the process from input to output; such as varying SR and SE throughout the data preparation process for variable selection and varying the resolution of locality (LR) in the calculation of local statistics (see Chapter 5). Four types of locality calculation are identified and one (adaptive moving window) is investigated in more detail (Fig. 5.1) and used in the prototype in Chapter 6. Such sensitivity analysis is one of six key typical analysis tasks associated with vPSA.

9.3.6 Visualising Sensitivities through gvPSA

Lastly the final research question (RQ4) – *Can data visualisation expose the sensitivities and uncertainties associated with spatial scale (resolution and extent) when comparing multivariate areal datasets?* – is answered through the creation (Section 6.3) and validation (Chapter 7) of the interactive visualisation prototype. The prototype was built to demonstrate the framework, in the context of energy-based geodemographic variable selection, with particular attention to varying the geographical aspect of SR, rather than SE or temporal or attribute based SR and SE. A visual exploration of

parameter sensitivity (Section 7.3) reveals clear sensitivities relating to varying the aggregation of SR at the standardise stage (StR) (see Fig. 6.17) as well as varying the SR of the calculation of locality (LR) (see Fig. 6.5). Fundamental changes to local patterns can be seen in the exploration of the sensitivity of geography through changing the number of neighbours (Section 7.3.1). The exploration of the sensitivity of StR at the global scale using the scale mosaic view identified 5 categories of sensitivity to multivariate relationships and pairwise correlation (Section 7.3.2). Where pairwise correlation mostly *strengthens* as StR increases, some variables are most sensitive to scale and values can *fluctuate*, *cross polarity* or even *weaken*. Due to limited data calculated at the LR level, further research is needed in order to investigate the combined relationship of StR and LR. In terms of transformation, the visual comparison shows clear differences in variables when transformed using the log scale, depending on the nature of the variable and the scale of the data (Section 7.3.3). Through the use of the prototype, varying the geographical representation from global to local in the variable selection process is shown to help inform the user of geographical variations within the variables when considering them for selection (Section 7.2.5). The consideration of locality within geodemographic variable selection may indeed influence future research in the area of geographically weighted geodemographics.

Visualising the effects of scale and the effects of locality have many possible extensions (explained in Section 7.7), yet the prototype demonstrates the utility of framework for the visualisation of multivariate data across scale with the inclusion of local geography. Both the framework and prototype were built and designed in the context of variable selection for geodemographics, where it is important to compare and understand the distribution, the correlation and the geography of the variables. These aspects helped to design the layout of the prototype, reordering options of P1 and the visual representations of histograms, scatterplots, map matrices and scale mosaics; however, both the model represented by the framework and its instantiation are transferable to other use cases as are the methods by which it was implemented. The broader applicability of the framework is demonstrated through the explanation of the two scenarios of Smart Home Analytics and Survey Response Modelling (Section 8.3) where multivariate data is necessary to compare across geography and scale, with the addition of temporal and attribute based SR and SE, instead of just spatial scale as demonstrated in the prototype.

9.3.7 The Design Science Approach

For this research, the new model was produced to tackle an unsolved design problem of visualising the complex parameter space of multivariate comparison with the inclusion of geography and scale. Design was crucial in the process and the instantiation of the model

was created through clear and justified design decisions. Design was seen as particularly important given the complexity of the parameter space. Some of the aspects of design demonstrated through the instantiation of the model are fundamental, such as the need for changing the visual representation and fluidity of the transitions between $V = \text{Multi}$ and $V = \text{Many}$, as discussed in Chapter 6. Although the artifacts produced in this research do not meet all of the seven guidelines for DSR by Hevner and Ram (2004), a number of them have been met; the framework and its instantiation can be classed as an “innovative purposeful artifact” (Guideline 1 ‘Design as an Artifact’) which is rooted in a specified problem domain (Guideline 2 ‘Problem Relevance’). The utility of the artifact must be shown in the specific problem (Guideline 3 ‘Design Evaluation’), which it achieves through the detailed investigation of Scenario 1 with the appropriateness of the design as well as the demonstration of its utility. Furthermore the second and third scenario form two additional demonstrations of its utility in context. To claim a DSR contribution, new knowledge must also be demonstrated. While it was known that variables were sensitive to scale, geography and transformation prior to this research, the design of the instantiation now allows these effects to be detected and explored, as demonstrated through the visual exploration of the parameter sensitivity in Section 7.3 where new knowledge about the geographical and scale variability of the energy variables is discovered.

DSR insists on the application of rigorous methods (Guideline 5 ‘Research Rigor’). While design consistency, coherence and definition were addressed in producing the instantiation and justifying the design decisions, there is more research needed in terms of optimising the methods; for example, there are limited (even offline) data processing solutions offered despite the multiple visualisation solutions demonstrated. The calculations were pre-processed and there was limited time to spend on optimisation of the R code or investigation of real-time possibilities. There is evidently a trade-off between reducing the scale of resolution (StR) or extent (StE) when including locality. As varying StE is not investigated in this instantiation, the use of locality was restricted by the use of detailed SR. The SR of LAD (326 regions) is the finest level of detail calculated for the instantiation at the full extent (England), yet a reduction of the extent would allow finer levels of SR to be analysed with locality. In the visual configuration of the instantiation there is also a balance between the number of L and V which are feasible to represent. Much of the precomputed locality data is not utilised in the instantiation as spatial aggregation is utilised in order to visualise L and V concurrently. The limitation of the configuration is also demonstrated when investigating the sensitivity of geography, scale and transformation in Section 7.3, where the visual components of the instantiation were re-arranged in order to have more pixel space for comparison. In terms to future research, an optimal balance is needed between processing power and visualisation power

resources. This is not seen as a drawback of the framework, but as a resource limitation in the computational context in which this design work was undertaken. There is great potential for continuing this research and improving the framework in future work. This is understood in DSR, as only when artifacts are studied is it possible to “*formalise the constructs, model and methods*” (March and Smith, 1995, pp.258).

The need for the framework was discovered as a research channel (Wood et al., 2014) in this project, and while it has been validated through the creation of the instantiation and demonstration of its utility, there are lots of open questions and many areas for continued research. Part of DSR is to ask questions of the problems and produce new knowledge: “*Each new program that is built is an experiment. It is posed as a question to nature and its behaviour offers clues to the answer*” (Newell and Simons, 1972 in March and Smith, 1995, pp.258). This is certainly true for this instantiation, in that the more it is demonstrated the more use-cases are found (see Section 8.4). One drawback of the current design is that it demonstrates a lot of information and the viewer is left overwhelmed by the possibilities. In order to reduce confusion, tailor the software, and meet DSR Guideline 7 (‘Communication of Research’) there is a three step course of action to follow:

1. Educate/Demonstrate the Problem
2. Educate/Training of Software
3. Software Redesign

Initially the problem needs to be demonstrated. Hevner and Ram (2004, pp.83) state that “*design-science research must be presented effectively*” to both technical and managerial researchers – who will study and extend them – and practitioners – who will implement them in their organisations. The research presented in this thesis, as well as feedback from presentations, reveals that even domain (energy, data and geographic) experts are often unaware of the benefit of visualisation or the breadth of complexities involved in multivariate comparison. Variable sensitivity across scale, geography and transformation need to be demonstrated through using the instantiation with clear examples in order to first educate the users and demonstrate the problem. Secondly, software training is needed in order to allow users to understand and master the software in its present state. This training can be evaluated as controlled experiments or simulations (Hevner and Ram, 2004) in order to establish user needs and to prioritise requirements for redesign. Finally the instantiation can be tailored to usable software solutions to meet user requirements, where additional or restricted functionality may be needed or a complete redesign for alternative scenarios.

In terms of creativity, the instantiation was not designed together with specific users and therefore is difficult to evaluate for creativity in the same context that the smart home

prototypes were evaluated (Section 3.5). Informal feedback from presenting the work at conferences, invited talks, workshops and to individuals in the field (of geodemographics and vis) has been particularly positive, with the views for scale mosaics, the asymmetrical matrix, and the mini simplified raster maps all being identified as *“interesting”*, *“novel”* or *“creative”* representations of the data. This links to Guideline 4 ‘Research Contributions’ where novelty of the artifact is crucial to claiming that an artifact is a contribution to research. In terms of visual design decisions, there are many features which draw on the initial aspirations from the Requirements Workshop, including: *“to see as much as possible on one screen”*, *“to show comparisons”*, *“to use known graphics”*, *“to slice and dice the data in different ways”*, *“see pretty, but precise data”* and to *“easily interact with the data”*. In general, this research demonstrates great potential for introducing aspects of creative design and design science into the visualisation design process.

9.4 Research Contributions

The primary contribution of the research is the theoretical framework for the investigation of scale and geography in multivariate comparison as described in Chapter 6. The model represented by the framework is demonstrated through the instantiation (Section 6.3) and its utility (Chapter 7 and 8). The visual designs produced for this thesis form another contribution demonstrating the visual possibilities of multivariate comparison across scale and geography. The results of the exploration and sensitivity analysis of energy variables (Chapter 7) and results of the research project with the analysts (Chapter 3) contribute to UK energy research. Analysis, designs and ideas in Chapter 4, 7 and 8 to improve the transparency of the geodemographic process contribute to ongoing research on open, domain-specific and local geodemographics. The inclusion of geography as an input into visual parameter space analysis (vPSA) to establish geo-visual PSA is a further contribution, explored through parameter sensitivity analysis in Chapter 7.

An additional contribution of the work is the use and investigation of creativity techniques and aspects of design science within visualisation design. Whilst these were not the main focus of the thesis, the creativity techniques were fundamental to the Requirements Workshop (Chapter 3) and the DSR approach provides a methodological grounding for the creation, application and evaluation of the framework (model) and prototype (instantiation). In addition to the model and instantiation a number of other DSR artifacts are presented in this research. New *methods* form new approaches to analysis related to visual design (scale mosaics) and techniques (creativity) and the potential for the new *constructs* of gvPSA and noClassification, where continued research is needed in order to study their utility. The creative requirements led not only

to the visualisation solutions for smart home analysis, but prompted the new channel of investigating geodemographics and the building of the framework and prototype in this context. The aspirations, ideas, barriers, discussions and storyboards from the initial workshop were all used to inspire and motivate the later stages of the research.

9.5 Research Limitations

There are a number of limitations to this research, which have been discussed throughout the thesis, but are summarised here for clarity.

The user stories for variable selection for geodemographics are implied from the literature, rather than working directly with geodemographic creators. This was due to the fact that there are only a limited number of geodemographic experts. One (Chris Gale the creator of OAC 2011 see Gale, 2014b) was informally consulted on his process, which confirmed the understanding of the processes and literature, as well as the added benefit of gaining knowledge about changes to the OAC methodology for OAC 2011 (which was still in process during this research). Continued research is needed with users in order to implement a usable and deployable solution for the creation of geodemographics (see Section 9.6).

The engagement with the energy analysts for the smart home project directly informed the smart home prototypes, where structured feedback and evaluation was sought. Their aspirations in this initial workshop inspired some of the later stages of the research when designing the prototype; however, the case study of geodemographics was seen as interesting for the customer and marketing department, rather than the participants themselves. Had the research focused on creating a visual tool to run and visualise the full geodemographic process (as outlined in Section 4.2) this could have been tested with relevant users in the company; however, the complexities and sensitivities of the process were seen as more important for academic research, therefore no user-based evaluation was necessary at this stage. The prototype was built to demonstrate the framework and was informed by the context of geodemographic variable selection. The specific designs are only evaluated internally. For continued research in the direction of variable selection for geodemographics, the consultation and evaluation with potential ‘middle-ground’ users (Ingram et al., 2010) (i.e. possible creators of bespoke geodemographics) would be necessary to evaluate whether the variable selection actually differs when including such visual representations of scale, local geography or transformations. The design research approach allows for the continuation of the process to ensure that the implemented solution is applicable in context.

The design of the prototype is also limited to the matrix view, as this was used in previous research and allowed for an overview of $V = \text{Many}$. Alternative space-efficient

visualisation techniques could be applied (e.g. McKenna et al., 2015), although further work would be needed to adapt these to multiple scales.

Reflecting on the connection to vPSA as described by Sedlmair et al. (2014). The framework, prototype design and evaluation in the context of variable selection for geodemographics only touches on the potential of gvPSA. The main vPSA aspect focused on in this research is the exploration of the sensitivity of the parameters of scale resolution (SR) and locality through varying N in Section 7.3. Sensitivity analysis is one of six vPSA tasks identified by Sedlmair et al. (2014). Further research is needed to explore the other tasks for effective gvPSA. Continued work also applies to alternative navigation strategies. The prototype and sensitivity analysis partly adopts two of the four vPSA navigation strategies of *global-to-local* and *informed trial and error*; however, continued research is needed in the area to evaluate their use for effective gvPSA. The main limitation of the work in the context of vPSA is that the *data flow model* is only partially implemented. The variable selection process is used for the input and output stages, whereas the full data flow model would see the classification as the final output in this context. Continued work on this area is discussed in the following section.

Whilst the prototype demonstrates applicability it also does not implement the full framework, as shown in Table 6.4. The applicability of the prototype is limited by the data available and subsequent data created during the process, i.e. the use of four scales, of which only two have local statistics at 3 varieties of N. This limits the comparison and evaluation of the sensitivity of scale and geography on the variables. With more time and resources this comparison could be more substantial. As geography and scale were given priority, transformation was only briefly investigated in this thesis by visualising the variables with or without the additional log scale transformation. All three aspects (scale, geography and transformation) need continued research to investigate these differences in more depth. The design and functionality of the prototype also limits the sensitivity analysis across many variables and additional visual representations were created for the parameter sensitivity analysis (Section 7.3). In addition, the nature of the local summaries mean that some summaries are based on small sample sizes. The investigation of geographical sensitivity in Section 7.3.1 used the LAD datasets with samples of 25, 50 and 100 rather than NUTS2, which had much smaller samples of 5, 10 and 15; however, the effects of small sample sizes has not been accounted for in the analysis at present. Further work is needed to investigate the nature of the local summaries and sample instability.

In terms of research questions, all were investigated but RQ1 was left relatively open as the industry is changing quickly. The question was investigated and partially answered through the choice of candidate variables used in the prototype and the evaluation of the

variables in Section 7.2. Overlapping variables were chosen to visualise in the prototype in order to allow for similar variables to be compared in the analysis and to make it possible to investigate known variable decisions in this context (with geographical variation and scale included). The requirements for the prototype could have been prioritised to visually investigate and explain why variables were changed from OAC 2001 to OAC 2011, and to choose selected variables to answer RQ1. This is partly addressed in Section 7.4; however, a thorough investigation was beyond the scope of the project.

Despite the limitations, the development of the prototype demonstrates how the framework can be employed in a particular context and the research provides a number of contributions to academia. Some useful extensions to the framework, designs (as discussed in detail in Section 7.7) and the research in general are outlined in the following section.

9.6 Future Work

Further work is necessary to continue to develop the research undertaken in this thesis. This includes extension of the sensitivity analysis. The effect of SE has not been tested in this research and nor have the other types of locality, e.g. fixed moving window or regular or irregular partitioning. Further investigation is necessary to determine the extent to which, and the circumstances under which, the variables with local variation affect the clustering results; for example the effect of ‘a strong global but a weak local’ or ‘a weak global but a strong local’ correlation. Continued development of the prototype in the context of geodemographics including many of the extensions outlined in Section 7.7 would allow users to be tested in order to evaluate whether the inclusion of geography, scale and transformations can better inform the variable decision process. Continued development of the visual process with the connection of the variable selection exploration stage to the other three stages (as outlined in Section 2.3.4) will allow users to also see the effect of the variables within the clustering process and the final results. Continued work in this context will see the vPSA’s *data flow model* (Sedlmair et al., 2014) realised in its entirety, with the energy geodemographic classification as the output of the model.

Continued work is also needed in defining the variables for classification and a further extension could see the use of smart home variables in an energy-based classification to produce more detailed profiles over time and include appliance use. This will combine the *who?* and *where?* of energy consumption to the *what?* and *when?*, to improve the understanding and potentially answer *why?*. Options for the combination of temporal as well as attribute data in the visualisation of the framework is discussed in the context of the two scenarios in Section 8.3 and the potential visual possibilities demonstrated in Section 6.2 and Section 8.1. Continued work to demonstrate the use of the framework in

this area would help to understand the affect of scale in greater depth. The continued development of the framework (model) through implementing prototypes (instantiations) for other scenarios, such as for the Smart Home Analytics and Survey Response Modelling scenarios (Section 8.3), will also allow the model to be tested more rigorously for its robustness, consistency and applicability (March and Smith, 1995).

Geodemographics simplify the characteristics of populations and allow for customer segmentation. The results of this research suggest that effective visualisation can be used as an alternative, especially in situations where data is plentiful (i.e. smart home data). Therefore a comparison of the use of specifically designed visualisation as an alternative to geodemographic classification needs to be tested *in situ* in order to evaluate the noClassification approach.

Finally, the continued research into the use of creativity techniques in data visualisation is seen as an important step to continue the research started in the preliminary stages of this thesis to combine the knowledge of the two domains. This could involve testing the creativity techniques used in this thesis, as well as additional ones deemed fit for purpose in a controlled study situation, to determine the creativity of the requirements and design outcomes.

9.7 Final Conclusions

“What is the future for household energy analysis?” and *“What value can be derived from energy consumption data through data analysis and visualisation?”* were two motivational questions posed when embarking on this PhD project. The investigation of the topic through literature, government reports, statistical and visual exploratory analysis and working with energy analysts during the smart home project, revealed that advanced data analysis and visualisation can not only improve knowledge in the industry, but benefit the shift towards a smarter and more efficient industry. Through investigating this topic, the research followed a typical DSR approach in that in-depth investigation of the problems related to the industry, combined with domain knowledge, triggered further questions and additional research as theories and artifacts were developed, justified and subsequently refined (Hevner and Ram, 2004). The research goal was adapted from an original proposal to create an energy-based classification, to the creation of a theoretical framework and prototype designs to enable others to understand the variable selection stage of creating profiles with greater ease. The resulting research stems from working closely with energy analysts, data modellers, creativity experts and visualisation designers.

The research incorporates innovative visualisation with design decisions referencing the visualisation literature. In addition to the overlap with vPSA, the inclusion of spatial scale and geography allows the work to be placed in the area of geovis, overlapping

visualisation with cartography, GIS and geographically inferred statistics. The ability to vary the locality in the framework and prototype encourages geography to be included in visual multivariate comparison and introduces possible uses for the visualisation designs. This includes aiding the explanation of global and local statistics and displaying the geographical variation of multiple variables. The research is useful not only to geodemographic creators, but may prove useful for geographers, statisticians, social scientists, data scientists or engineers. The framework and visualisation examples can be used for improved variable selection for geodemographic classification (and other forms of classification or clustering involving variables with geographical variation) as well as other multivariate comparison purposes where there is a need for understanding the sensitivities of data scale and/or geography. Such examples include the two additional scenarios where the framework is shown to be applicable to the exploration of smart home datasets to improve knowledge on energy consumer habits and behaviour, as well as aiding the identification of variables to identify nonresponse bias in surveys.

In general, the research in this thesis has investigated many possibilities for visualising household energy consumer characteristics. It identifies a need for energy-based profiling, candidate variables to use for profiling and investigates visual designs for enabling the complete profiling process to be more transparent. The framework for visualising multivariate data is the fundamental contribution of the research which allows the sensitivities and complexities of the data variables to be investigated. Through the design science approach, the work undertaken demonstrates that (creative) visualisation can be useful in providing rich and meaningful perspectives on complex datasets and that the visualisation of multivariate data across multiple scales with the inclusion of geographical variation is not only possible but important.



Academic Impact

The following pages contain peer-reviewed (A.1-A.4) publications and short papers written during the duration of the PhD research for international and national conference presentations and posters. The work was also presented at the workshop on GeoVisual Analytics: Interactivity, Dynamics, and Scale, at the giScience 2014 conference, 23 Sept 2014, Vienna, Austria (same paper as A.2), the North American Cartographic Information Society (NACIS) annual meeting (A.5 abstract only. Includes the presentation slides for context - as these designs are referenced in the thesis) and a PhD symposium on Household Energy Consumption at Birmingham University (A.6). The papers are ordered by type – peer-reviewed article, short papers and other presentation – then in reverse chronological order. All appendices A.1 - A.6 provide substantial background, supplementary content and reasoning for undertaking the research and show progressive research findings. Additional citation details for each of these appendices are available at the start of the thesis.

A.1 Article: TVCG 2013, 19(12), pp.2516-2525

Preprint paper for IEEE Transactions on Visualization and Computer Graphics
(Proceedings of Information Visualization 2013) 19(12), pp. 2516-2525
Available here: <http://openaccess.city.ac.uk/2618/>

Creative User-Centered Visualization Design for Energy Analysts and Modelers

Sarah Goodwin, Jason Dykes, Sara Jones, Iain Dillingham, Graham Dove, Alison Duffy,
Alexander Kachkaev, Aidan Slingsby, Jo Wood, *Member, IEEE*

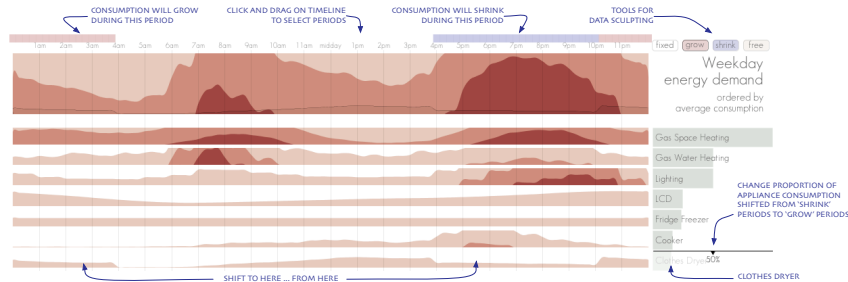


Fig. 1. *Demand Horizons* show modeled weekday energy demand over 24 hours amongst high consumption domestic appliances. *Data Sculpting* allows us to shift consumption interactively by 'moulding' the horizons to explore 'what if?' scenarios. For example, here fifty percent of 'Clothes Dryer' consumption is shifted from the evening peak to a period when overall demand is lower.

Abstract— We enhance a user-centered design process with techniques that deliberately promote creativity to identify opportunities for the visualization of data generated by a major energy supplier. Visualization prototypes developed in this way prove effective in a situation whereby data sets are largely unknown and requirements open – enabling successful exploration of possibilities for visualization in Smart Home data analysis. The process gives rise to novel designs and design metaphors including *data sculpting*. It suggests: that the deliberate use of creativity techniques with data stakeholders is likely to contribute to successful, novel and effective solutions; that being explicit about creativity may contribute to designers developing creative solutions; that using creativity techniques early in the design process may result in a creative approach persisting throughout the process. The work constitutes the first systematic visualization design for a data rich source that will be increasingly important to energy suppliers and consumers as Smart Meter technology is widely deployed. It is novel in explicitly employing creativity techniques at the requirements stage of visualization design and development, paving the way for further use and study of creativity methods in visualization design.

Index Terms—Creativity techniques, user-centered design, data visualization, smart home, energy consumption.

1 INTRODUCTION

These are exciting times for utility companies and their energy analysts – the energy domain is data rich and globally significant. Energy analysts and modelers are now striving to effectively use the volumes of data from emerging Smart Home technologies to understand consumer behavior, conserve energy and manage supply and demand. Data visualization can offer great potential in this domain, but developing appropriate solutions presents considerable challenges, since the nature of the data are relatively unknown and the needs of energy data analysts and modelers are not yet well understood. The design brief is therefore essentially open-ended.

- Sarah Goodwin, Jason Dykes, Aidan Slingsby, Jo Wood, Iain Dillingham, Alexander Kachkaev are with the *giCentre*, City University London. E-mail: {Sarah.Goodwin.1, J.Dykes, A.Slingsby, J.D.Wood, Iain.Dillingham.1, Alexander.Kachkaev.1}@city.ac.uk
- Sara Jones, Graham Dove and Alison Duffy are with the *Centre for Creativity in Professional Practice*, City University London {S.V.Jones, Graham.Dove.1}@city.ac.uk, alison@perspectiv.co.uk

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 27 September 2013.

For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Participatory approaches to user-centered design, in which users and other stakeholders are involved in co-creating requirements and designs for interactive systems can lead to solutions that are more useful and usable [35]. We have successfully used human-centered approaches in the design of visualization solutions before and have documented these in detail [27]. However, the role of *creativity* in these approaches has as yet been only implicit. Over the last decade some fields of interactive systems development have increasingly focussed on introducing elements of deliberate creativity into participatory user-centered design processes. The aim here is to enable all participants (users, designers and other stakeholders) to contribute to the exploration of new fields and the generation of requirements and design ideas for novel and useful systems [1, 6, 53]. Establishing requirements can be considered a fundamentally creative process whereby requirements analysts and stakeholders work collaboratively to generate ideas for software systems [29, 30, 32]. Indeed, Robertson [42] regards requirements analysts as inventors who bring about innovative change in designs to establish advantage. Techniques for deliberately introducing creativity into the process of user-centered design can be used effectively in this context. For example, Schmid [46] used creativity triggers [42] to help workshop participants invent requirements, whilst co-creation [45] and creativity workshops [24, 31] have been shown to be effective in generating novel requirements.

Here, we report on work in which we augment a user-centered ap-

proach to design with techniques for deliberately stimulating creative thinking when establishing context of use and developing requirements. We do so in the context of an investigation into ways in which a major energy supplier could use visualization to derive value from data that will become available following the wider adoption of Smart Home technology, by producing a series of prototypes to establish visualization possibilities. We evaluate the prototypes in terms of appropriateness, novelty and surprise and conclude that the creative impetus to our design activity had a long-term effect, contributing to designs that were found to be effective, informative and novel and a process in which creativity flourished. We offer a series of contributions that may be useful in energy visualization and beyond, namely:

- i. a *creative design case study* where a user-centered process is augmented with means of deliberately stimulating creative thinking;
- ii. *techniques* for the visualization of a new data source, including methods that contain some novelty, that may be transferable as data of this type becomes more common and voluminous;
- iii. evaluation of *creativity methods* in an applied context to support the contention that deliberately stimulating creative thinking can result in designs that are novel and useful – especially in the context of open requirements in problem-driven visualization.

2 APPLIED CONTEXT

Smart Meter technology enables energy consumption to be recorded for multiple appliances within the home at frequent intervals. Data are reported back to both energy supplier and consumer enabling near real-time feedback on energy use. The European Commission recommends all member states adopt intelligent meter technology with the majority to be fully equipped by 2020 [13]. The installation of Smart Meters forms a major component of the shift from passive electricity supply to ‘Smart Grids’, which use digital technologies to manage the regulation of energy demand and production, allow for flexible tariffs and provide the potential to communicate directly with Smart Homes or appliances [13]. Advances in Smart Meter technologies are consequently becoming increasingly important to both energy suppliers and consumers, whilst data yielded from these new technologies is increasing the volume and value of data available to the industry exponentially [44]. Energy data analysts and modelers are beginning to investigate opportunities to utilize the emerging data to understand consumption trends and consumer behavior [14] and to manage supply and demand effectively through optimization and flexible tariffs [4].

Data visualization and visual analytics offer real opportunities for the analysis of Smart Home data both for the energy supplier and the consumer. On the consumer side, energy use information is reported through a Smart Energy monitor. While this is seen as beneficial in comparison to the traditional energy bill [18] less intrusive forms of consumption awareness are now being investigated [43]. Visualization solutions to enable the energy industry to gain valuable insight into customer habits, identify areas where consumption can be reduced and effectively manage supply and demand levels are, however, scarcely investigated in the literature. The benefits of using visualization to study aggregated household energy use to discover patterns and trends have been highlighted [12], however the data are based on diary entries rather than volumes of frequent automated recordings.

Our research with data analysts from a major UK energy supplier begins to investigate the benefits that data visualization can bring to derive value from the data emerging from Smart Home technologies and opens up opportunities for further research. It uses the two sources of Smart Home data currently available: *live data* from a Smart Home trial and *modeled data* simulating future scenarios. The live data contains electricity and gas consumption for all appliances (e.g. refrigeration unit or television set) as named by owners of a test-bed of 130 properties participating in a Smart Home trial. The data set consists of more than 18 million recordings taken over a 14 month period. It has challenging characteristics: timings are irregular; frequency of recordings varies significantly – from minutes to days; the sample of households is small (the UK contained 24.6 million households in 2012), self-selecting and biased in terms of geography and demographics.

Householders are also inconsistent in the appliances they monitor. The model [17] uses a separate source of detailed consumption data [56] to generate appliance-based energy usage scenarios for any number of households at 15-minute intervals over a given period of time.

Both sources contain numeric information for individual households (modeled or trial participant), such as total electricity consumption, consumption by individual appliance or outside temperature, along with the time of the recording. Derived values (average, max, min, count, standard deviation) are calculated in both cases by period of time (hour, day, week etc.) and by grouping categories (such as appliance type). The model can generate large volumes of data in this form with optimized outputs simulating the shifting and reduction of demand over time. Different outputs reflecting weekday and weekend activity are also available. Daily and seasonal variations in consumption and standby options are modeled with some sophistication for certain appliances. Outputs are somewhat limited however, in that appliance use and distribution of appliances to households are determined probabilistically [17] and so may not reflect real ownership or typical household usage patterns. Appliance co-ownership relationships are therefore not realistic and neither household demographics nor geographical location are accounted for in the simulation.

3 CREATIVE DESIGN PROCESS

Our design process for exploring the possibilities for data visualization within Smart Home data analysis followed an established user-centered approach [25, 27]. However, we augmented this by applying a number of creativity techniques [24, 29, 31, 37] early on in the process. Our aim here was to see whether we could tap into the latent creativity of our target users – the energy analysts – as well as that of the design team. While designers, of visualizations and other artefacts, may be used to developing creative responses to problems or design briefs, their customers, users, and other stakeholders may not be. We have previously employed such deliberate creativity techniques with air traffic controllers [31] and the police [38], who have not been accustomed to making creative contributions to design. Through the use of techniques such as those described below, they have, in each case, been able to generate requirements and design ideas for new interactive systems that were considered both novel and useful. Here we apply these methods alongside our established means of encouraging data owners to engage actively in visualization design and development [25, 26, 27, 41, 49]. The process is summarized in Fig. 2 with the creativity techniques being inserted in the early stages with the intention of introducing a creative climate that we hoped would persist.

3.1 Creative Requirements Workshop

Creativity techniques for use in our *Requirements Workshop* were developed through two internal pilot sessions. Techniques from methodologies such as creative problem solving (CPS) [37] and Syntectics [16] were considered and additional literature reporting similar techniques was consulted [22, 34]. These included: aspirational thinking, analogical reasoning, metaphor, constraint removal, storyboarding and random combination. We tried methods out internally and adopted the techniques that were thought to be most practicable and potentially useful whilst rejecting some that might constrain – such as building a priority list or listing ideas based on their complexity. We augmented others, such as an established “*I wish*” exercise for wishful thinking [33] with prompts specific to the visualization context – “*I would like to see*”. The methods were refined in collaboration with a professional creative facilitator, who coordinated the *Requirements Workshop*.

As well as tailoring the creativity techniques, we also paid careful attention to our choice of venue, as the physical environment in which activities are carried out can have a significant impact on the creative climate [11, 23]. We therefore chose to carry out the workshop in a quiet, light, neutral venue, away from the participants’ normal places of work, with plenty of space and ample refreshment. The day long event was attended by five Smart Home energy analysts, who work together on a regular basis. They are often involved in thinking of new ideas and possibilities for Smart Home technologies, however, their

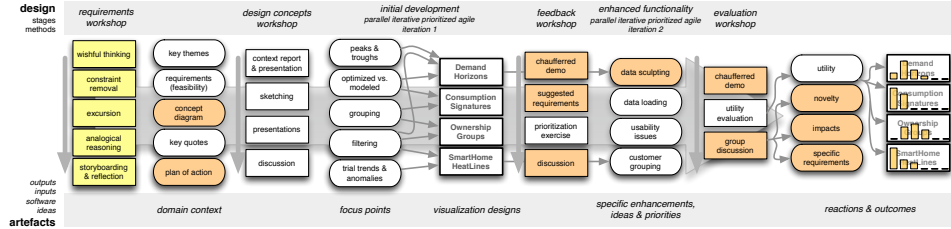


Fig. 2. The design process. Rectangles are techniques, those with thick edges represent software prototypes. Concepts are round edged. Arrows show direct links between concepts and prototypes. Other links are implicit and less direct. Yellow indicates deliberate creativity mechanisms. Orange highlights processes and concepts in which creativity amongst analysts was strong. Prototype utility is reported in detail in section 5.

knowledge of the new data sets available to them and the opportunities offered by data visualization were limited.

We began with some warm up activities. These included a playful introduction that encouraged participation and trust-building and introduced some analogical thinking by asking all participants “if you were to describe yourself as an animal, what would you be?” Some statements and quotations that emphasized creativity and exploration were also shared – for example, Albert Einstein’s widely reported view that: “if at first, the idea is not absurd, then there is no hope for it.”

3.1.1 Wishful Thinking

The first creativity technique employed in the main part of the workshop was wishful thinking, in which the energy analysts were asked to think about aspirations for the Smart Home programme. We captured visualization specific ‘opportunity statements’ [23] by asking: ‘What would you like to know?’, ‘What would you like to be able to do?’ and ‘What would you like to see?’ Participants worked individually on Post-it notes in a brainstorming [36] exercise, then read their answers out to the group and placed them on flip-charts. We then asked the participants to form small groups and each was tasked with selecting the Post-it in which they were most interested. To push them further in their thinking, the analysts were asked to consider ‘What next?’ and further aspirations were recorded (again on Post-its) assuming the chosen aspiration(s) had been achieved. The process continued until ideas were exhausted and some initial requirements had been teased out, revealing some of the types of innovation in which participants were interested.

3.1.2 Constraint Removal

After coffee, participants built upon this forward thinking with a constraint removal activity [24] in which barriers were transformed into a positive resource through which to create new ideas. Our energy analysts were first asked why the aspirations captured on Post-its had not yet been achieved. Once constraints were identified analysts were then asked for creative ideas about what would be possible if the barriers were removed to see whether ideas would develop further. A rapid flow of constraints resulted – from hardware technical issues, to people leading complicated lives and being difficult to understand, limited knowledge about Smart Homes, a lack of customer trust, limited time, resources and expertise as well as conflicting business priorities. ‘Removing’ some of these constraints unlocked a number of ideas about moving forward: in particular about improving and expanding the product, gaining the trust of customers and the energy industry and deriving value and knowledge from the live Smart Home data source.

3.1.3 Lunchtime Excursion

Lunch was held in an adjoining building during a lengthy break. Participants were asked to use this time to find something that had a connection (however abstract) with the Smart Home programme. This was based on the idea of an ‘Imagery Trek’ in CPS [36] or ‘Excursion’ in Syntectics [16]. Both are techniques that can help develop highly novel

or unexpected ideas and assist participants in refining or elaborating their ideas through ‘mental stretching’ [23]. The idea is that participants remove themselves from a task, take a mental or physical journey to seek images or stimuli and then bring these back to make connections with the task. Participants returned from their excursion with all sorts of artefacts including photos of a painting and the view from the lunch room and a copy of Dickens’ ‘Great Expectations’. This activity set the scene for the subsequent analogical reasoning task.

3.1.4 Visualization Awareness using Analogical Reasoning

The analogical reasoning task was an extension of the ‘Visualization Awareness’ activity that is central to our existing human-centered visualization design process [10, 25]. Here, however, we began by specifically explaining analogical reasoning and giving examples. We then asked the analysts to find analogies applicable to Smart Home visualization as they engaged in an otherwise relatively passive visual experience that introduced visualization examples by theme. Participants were given time to consider any aspects of the examples (data, layout, interactions, colors, aesthetic) that sparked a connection with the thinking that had occurred during the morning sessions. Reactions were again written on Post-its, and some of the participants created mind-maps to link the different visualizations to their ideas. In total ten analogical ideas arose while watching the visualization demos, including an idea to show wasted energy flows that was sparked by an animated visualization of millions of bike journeys [55] and an idea for using bubbles of energy consumption increasing and decreasing as used in the home, inspired by *Empires Decline – Revisited* [7]. Design requirements identified during the exercise included the need to filter, group and compare data such as by appliance type, temperature, user demographics, time and geography to understand consumption variability. Design elements identified as important included: ‘everything in 3 clicks’, ‘beautiful’, ‘engaging’ and ‘simplicity’.

This activity took longer than planned, largely due to the large number of wide-ranging and increasingly ambitious ideas that surfaced. The session ended with a highly creative *Plan of Action* envisaged for the focus of Smart Home data analysis involving a three stage process to which we could make an important contribution, namely:

1. discover – find out where energy is used;
2. displace consumption – change behavior and control devices;
3. reduce energy production – specifically by the amount needed to close a power station (power plant).

3.1.5 Storyboarding

We have used storyboarding [3] previously in creative requirements workshops in other domains [29, 30, 31, 32] to draw together and prioritize the ideas generated. Here, pairs of participants used a comic strip template, writing materials and hard copies of the various visualization awareness examples to generate artefacts (sketches and collages) showing how the ideas generated during the day might be used in practice by imagining ‘a day in the life of an energy analyst’.

Table 1. *Wishful Thinking* revealed in 'Know/Do/See' and 'What next?'

Activity	Aspiration Topic	Total	Feasible
Know	Customers Habits	10	5
Know	Appliance Consumption	6	6
Know	The Value of the Data	2	2
Know	Visualization Design	2	2
Do	Improve Customer's Experience	5	2
Do	Manage Energy Demand	3	3
Do	Advance the Technology	3	0
See	Data Analysis & Visualization	8	6
See	New Products and Services	1	1
—	—	—	—
Next?	Change Customer Behavior & Improve Life	5	0
Next?	Improve & Expand the Product	6	0
Next?	Understand Customers Habits	3	2
Next?	Gain Trust & Increase Customers	5	0
Next?	Educate Energy Industry & Manage Demand	5	1

Key themes that emerged from the storyboards included the need for greater understanding of consumers' habits and the desire to understand customer behavior by grouping and comparing relevant data.

3.1.6 Reflection

To round off the workshop, participants were asked what they knew at the end of the workshop that they hadn't known at the outset. Their responses at this point were very positive, both in regard to the possibility of developing appropriate visualizations "It's amazing how many techniques are applicable to energy" and in regard to the workshop itself "I understand more about the large scope of possibilities."

Overall the outcomes from the day's activities allowed us to identify five key themes that can be seen as important to the continuation of the Smart Home programme: *Analyze the Data*: to understand more about customers' energy habits and appliance consumption; *Develop Knowledge*: to start to prove / disprove myths and theories of energy saving and behaviors; *Communicate and Engage*: within the business, and with industry and the general public to manage demand and change behaviors; *Build Trust*: in the company and the products; *Improve and Expand Smart Products*: beyond energy to improving comfort and security. The first of these themes links directly with the first stage of the *Plan of Action*: *discover – find out where energy is used* (see end of 3.1.4), a key objective in which visualization can play an important role. Improving the understanding of customer and appliance consumption will also help pave the way to targeting some of these other themes and reaching the second and third stages in the *Plan of Action*.

The wishful thinking exercise generated 64 aspirations and opportunities of broad scope as shown through their grouping into topics (Table 1). We identified 30 of these as feasible for data visualization solutions in terms of the expertise, data and other resources available.

These key themes and feasible aspirations were reported to designers and developers in the team along with other artefacts to help them gain a broader understanding of the analysts' needs and identify where and how effective data visualization design might be beneficial, as described below.

3.2 Design Concepts Workshop: Development Iteration 1

Development took place over a one month period with two iterations using a rapid agile approach. Within each iteration features were prioritized using the MoSCoW technique [2] with frequent meetings between designers and developers in the team to re-prioritize and discuss design decisions in light of requirements.

The first iteration began at a half-day *Design Concepts Workshop* that brought together seven visualization designers and developers (all are co-authors) many of whom had limited background knowledge of the energy industry. We began the session by presenting and sharing the domain knowledge as well as the key themes and ideas from the *Requirements Workshop*. Contextual information including the 3 stage *Plan of Action*, key themes, feasible aspirations, design requirements,

mind-maps and a concept diagram generated in part from these, storyboards and some direct quotes were introduced and then pinned to the walls of the room in order to prompt movement, discussion and idea generation amongst designers. The two energy data sets were also introduced and their structure, provenance and limitations discussed.

Working in pairs we generated ideas, developed sketches and reported back to the group with reference to the requirements that the idea was targeting. This enabled us to derive visualization focus points – abstract combinations of task, data and design that form a basis for ongoing development: show *peaks and troughs* in daily demand to understand when different appliances are used; *compare modeled to optimized solutions* to see whether shifting consumption could help demand management; *group and filter* consumption by appliance and types of appliance across time to identify patterns in user behavior; and, identify *trends and anomalies* in the Smart Home trial data.

These focus points were further developed during the workshop and through subsequent activity into four prototype visualization designs. These addressed generic aspirations from the wishful thinking exercise, such as: "to know how to show the business stakeholders the data in an engaging way," "to find typical patterns and make predictions," "to know where energy is going" and "to 'slice and dice' the data," as well as specific aspirations and questions as follows:

Demand Horizons: highlight the peaks and troughs in the modeled hourly energy demand during typical (weekend and week) days and show how each appliance contributes – "to know what an 'average home' does with their energy" and "to better understand how different appliances contribute to the peaks in energy demand throughout the day."

Consumption Signatures: show how each appliance has a different signature over time-of-day and day-of-week in the modeled data by visualizing large amounts of energy consumption data in comparable form on one screen – "how can we visualize large amounts of energy consumption data on one screen?" and "can we compare the energy consumption signature of appliances or groups of appliances?"

Ownership Groups: group appliances in the modeled data by ownership, time of use and average consumption – "to know how lifestyle links to energy demand" and "to better understand how the data relates to the users."

Smart Home HeatLines: a per-household representation of the live Smart Home trial data to identify patterns and anomalies – "how to visualize all the data from the Smart Home trial to understand the usefulness of the data?"

Developing designs in parallel enabled us to address multiple focus points concurrently, present alternative techniques of potential value to the domain experts and use an established means of generating high quality and diverse outputs [9]. It also offered plenty of 'breadth' in terms of enabling us to explore opportunities for ongoing creativity.

3.3 Feedback Workshop: Development Iteration 2

Following the first development iteration a number of enhancement possibilities were suggested and associated effort estimated for each. These possible enhancements were the focus of a *Feedback Workshop*, involving the four analysts who had taken part in the *Requirements Workshop*, and four others from related departments in the same organization. We presented the aspirations gathered from the *Requirements Workshop*, reflected on how we had formulated these into focus points and demonstrated our initial designs by chauffeuring the visualization prototypes in an engaging and increasingly interactive visualization session held at the company's Smart Home test house.

Initial reactions, new ideas and other feedback were recorded for each design prototype. Our proposed enhancements and any suggestions identified during the session were then prioritized by the group.

After the session, enhancements for each prototype were considered through a systematic re-prioritization process in terms of development complexity, time available, novelty of idea and priority through an agile procedure for planning estimation [5]. The enhancement prioritization progress is itemized in Table 2 and a number of key new features for each prototype were implemented as described below.

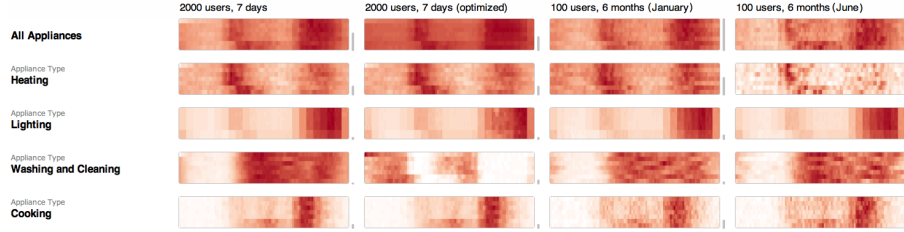


Fig. 3. *Consumption Signatures* allows modeled data to be loaded (columns) and reordered so that the weekly consumption patterns of appliances can be compared. Various coloring options scale sequential schemes by selected row, column or cell and allow diverging schemes to emphasize difference from selected items. Patterns in daily (*Lighting*), seasonal (*Heating*) and modeled (*Washing and Cleaning*) data are clear as are weekend differences (bottom two rows of each cell) such as the delay in the morning heating peak and more cooking during daytime at weekends.

Table 2. Prototype Enhancements

Prototype Name	Presented	User-Suggested	Implemented
<i>Demand Horizons</i>	11	7	6
<i>Consumption Signatures</i>	7	5	10
<i>Ownership Groups</i>	10	3	8
<i>Smart Home HeatLines</i>	10	3	6

4 RESULTS: VISUALIZATION PROTOTYPES

The four prototypes were designed and developed with complimentary characteristics to explore different tasks, data and designs – as characterized by the focus points (section 3.2). The features are described below with detail of specific interactions explained and demonstrated in the supplementary video.

4.1 Modeled Data

Two of the prototypes used hourly consumption data modeled for 2000 households over a period of 30 days, with different average hourly rates calculated for households at weekdays and weekends

Demand Horizons (Fig. 1) uses horizon charts [20] to show aggregated and appliance-based energy demand during a typical 24 hour period. Horizon charts can be instantly switched to area graphs in order to aid understanding. Animated transitions [21] highlight the differences in consumption between typical days during the week and weekend. Appliances can be re-ordered according to their contribution to the total, morning or evening peaks and individual appliance charts can be added or removed for detailed investigation of the differences in demand between appliances and their effect on overall consumption. Several amendments were implemented in the second development iteration, including quick switching between gas and electricity appliances. In particular, a new feature was created in order to allow demand to be modified directly through the metaphor of *data sculpting*. This allows peaks to be flattened through the interface in two ways: the overall consumption of any appliance can be interactively varied to simulate improved efficiency; consumption can be time-shifted, using the *grow*, *shrink*, *fix* or *free* buttons, to simulate change in behavior (see Fig. 1 and video). *Ownership Groups* (as shown in the supplementary video) consists of a bar chart linked to a set of Tufte’s [50] redesigned Tukey box plots [51]. Bars representing each appliance are sized by the number of households that own at least one of each. Bars can be re-ordered to show the appliances by proportion or alphabetically. The box plots show average hourly consumption of households. Upon selection of a particular appliance these are updated to show the average consumption of the households owning this appliance. Design enhancements implemented after the *Feedback Workshop* included new selection mechanisms and three additional means of ordering – by appliance type, subtype and total power/load on the grid. Alternative views related to co-ownership of appliances were also investigated.

Consumption Signatures (Fig. 3) visualizes the model’s highest resolution data, with records at 15 minute intervals aggregated according to time of day and day of week. Multiple outputs can be structured in to this weekly *signature* for comparison, including a six month simulation to show seasonal variation and a one week simulation with two algorithmically optimized alternatives. Multiple derived values (such as minimum, maximum and average consumption) were abstracted from the model outputs and households were sampled in the case of large data sets to ensure rapid responses. Calendar views [52, 54] visualize weekly consumption: seven rows relate to days of the week, with 96 columns representing each 15 minute period of the day. Signatures are positioned in a matrix of small multiples in which data sets (columns) and appliances or groups of appliances (rows) are juxtaposed for comparison [15]. The signatures are colored according to their values with two alternative schemes: a sequential scheme represents absolute values and a diverging scheme [19] shows the numerical difference between each signature and a selected item: a column (data set); row (appliance); cell (particular signature) or pixel (individual value). During the second development iteration the need to rescale the legend to the ‘best fit’ for each signature was identified and implemented.

4.2 Smart Home Trial Data

Smart Home HeatLines (Fig. 4) represents the raw live data from the Smart Home trial. Individual households are represented as rows of values varying over time. Summaries (count, average, maximum and minimum) are calculated by household for each variable for particular time periods. Further data abstraction is available in real time as the temporal kernel can be interactively re-sized to aid pattern identification and avoid distortion due to inconsistencies in collection times. Sequential color schemes [19] are used to represent values, with a line graph to aid in the identification and interpretation of patterns and trends for any selected household. The summary statistic, source (electricity, gas or appliance) and time period (total and weekly or daily averages) can be varied interactively. Households (rows) can be re-ordered by value at a particular time period. Grouping by demographic type, sorting by similarity of profile and a map to show animated geographical variations over time were added during the second development iteration – as shown in the supplementary video.

5 RESULTS: VALIDITY AND CREATIVITY

Reflecting on both the visualization design evaluation literature [47] and methods for evaluating creativity [8, 28] we constructed a structured process to determine the extent to which both the visualization prototypes themselves and the design process through which they were generated were seen as both valid and creative (Table 3).

The extent to which the outputs of our process were themselves viewed as creative was a particularly important indicator of how successful we had been in our introduction of techniques for deliberately

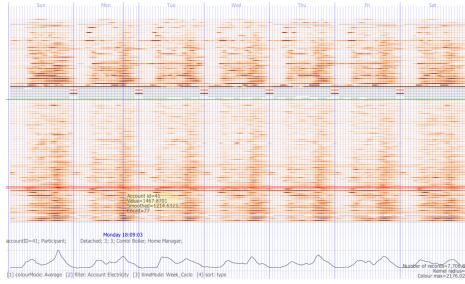


Fig. 4. *SmartHome HeatLines*: visualizes Smart Home trial data per household by time. Here, data are aggregated to show average weekly electricity consumption, with households ordered (top to bottom) by type of participant and consumption on Monday at 6pm.

stimulating creative thinking into the design process. A review by Dean *et al.* [8] reveals that most authors evaluate creative outputs through some combination of the dimensions of appropriateness, novelty and surprise. Our evaluation was therefore structured in this way, with questionnaires, a structured group discussion, and subsequent analysis of responses. The objective was to gather analysts' views of the *appropriateness* of the designs, in terms of whether or not they satisfied relevant requirements, their *novelty*, in relation to the analysts' previous experience, and the *surprise* that they engendered.

We conducted an *Evaluation Workshop* with four of the five energy analysts who participated in the *Requirements Workshop* at the Smart Home test house. We began by presenting the four prototypes and demonstrating the enhanced functionality that had been added during the second development iteration through (increasingly analyst directed) chauffeuring, linking this to specific requirements and feedback. Chauffeuring was deemed appropriate as a rapid means of getting analysts to use the software to access the data and as we were not evaluating the usability of the prototypes but rather the value of the approaches developed in regards to established opportunities.

After each demonstration analysts evaluated the appropriateness, or utility, of each prototype by completing a questionnaire that asked them to assess the extent to which various relevant requirements were satisfied by the prototype by rating strength of agreement on a six point scale ranging from strongly agree (1) to strongly disagree (6). Due to the small numbers of prototypes and participants involved in the study, it was not appropriate to attempt any quantitative evaluation of the novelty or surprise factors of the prototypes, and we therefore adopted a qualitative approach to evaluating these aspects. Thus the *Evaluation Workshop* ended with a structured group discussion where the prototypes were again used through directed chauffeuring on a shared screen to prompt discussion relating to the novelty of each design, and the surprise they engendered.

Our aim in evaluating the creative user-centered process through which the designs were developed was to gain some initial insights into the extent to which it could be seen as being effective and creative, and the impacts this may have had on designers and other stakeholders, as well as on the prototypes that were developed. We relied predominantly on the reflections of our experienced design team, informed by inputs from other stakeholders during the structured group discussions (see section 5.3), as documented in section 7.

5.1 Appropriateness of The Prototypes

Responses to the questionnaires reveal that 3 of the 4 prototypes score highly for meeting the needs of the energy analysts as expressed during the *Requirements Workshop* – responses tending to the left in Fig. 5.

Demand Horizons returned a modal score of 2 for the questionnaire responses, and the energy analysts thought of many uses for the tech-

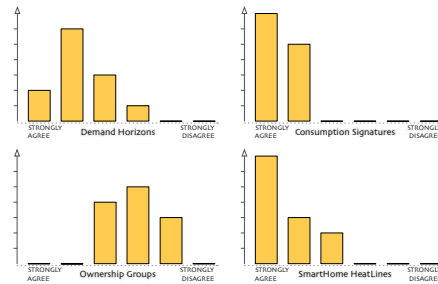


Fig. 5. Responses to the Prototype *Appropriateness* Questionnaire. Strong agreement (1) with positive statements about utility in light of requirements to the left, strong disagreement (6) to the right.

Table 3. Evaluation Process

Considering	Evaluating	Method
The Prototypes	Appropriateness	Questionnaire
The Prototypes	Novelty	Structured Group Discussion
The Prototypes	Surprise	Structured Group Discussion
The Design Process	Validity & Effect	Structured Group Discussion
The Design Process	Creativity	Reflection by Designers

nique, some of which were beyond the initial remit: “it starts to become an interesting customer’s view.” The analysts found the design particularly appealing and engaged especially with the *data sculpting* feature, which is discussed in more detail in section 6.

Consumption Signatures scored 1s and 2s in the questionnaire (signifying *strong agreement* or *agreement* that requirements were satisfied). The energy analysts were excited and fascinated by this application. It was seen as “*very powerful and very useful*,” highlighted as being a particularly intuitive design that allowed analysts to gain insights quickly: “*you could spend months searching the data for insights but this just points you straight at it.*” It was also seen as an excellent knowledge building tool: “*I could imagine ... just taking a week off and just letting your curiosity dive in and out.*”

Ownership Groups scored 3s – 5s in the questionnaire and was the only prototype not seen as immediately useful by the analysts. While the questions being asked were notably valid and useful to the industry: “*just knowing what people have allows you to size up the market,*” the modeled data does not group appliances with users in realistic ways. This lack of validity in our data limited opportunities for insight and thus utility. The slick and elegant design, whilst meeting the criteria gathered from the *Requirements Workshop*, was in part also deemed inappropriate – the Tufte [50] style box plots being unpopular.

Showing the live trial data through *Smart Home HeatLines* caused particular excitement and engagement. All scores were between 1 and 3 with a mode of 1 indicating that it was considered highly relevant to the analysts’ needs. The tool was deemed appropriate for “*a very wide user base*” in fact “*anyone interested in gaining insight from energy consumption data.*” The focus group discussion also revealed that it could improve communication of the Smart Home project amongst colleagues: “*we could be there for days, sharing it with other people.*” The value of exposing the analysts to the trial data in this way was explicit: “*this would be invaluable in starting to prove that some of these electronic [Smart Home technology] approaches work.*”

Alongside our evaluation by energy analysts, we also asked the energy modelers, who had generated the data on which three of the prototypes were based, to informally evaluate our prototypes. We engaged with them throughout the development process and found that they considered all four prototypes very appropriate to the needs of the en-

ergy industry and in particular to the needs of a modeler: “*The way you solve a problem is by doing some visualization in your mind and these tools help you greatly to facilitate that.*”

5.2 Prototype Novelty and Surprise

The four design prototypes in general were described by one of the analysts as “*creative approaches which show us the density, variability and value of our data.*” The techniques used were “*very different*” and new to the analysts: “*the methodologies would not have come out of my head.*” Overall the designs were deemed novel and valuable: “*you have brought something that we couldn't have thought of ... and the [Smart Home] project will be better for it.*”

Novelty and surprise were expressed in reactions to *Smart Home HeatLines* during the *Evaluation Workshop*: “*I think this is brilliant*”; as well as after reflection in the *Evaluation Workshop*: “*it gives us a whole new way of analysing people, “18 million data points! [It] is just impossible for us to get our head around the real value that is contained in that” and “I did not realize how diverse the different profiles were.*” The prototypes visualizing the less familiar modeled data also resulted in expressions of surprise and evidence of novelty. The heat mapping in *Consumption Signatures* can not be termed novel as a technique, but the sheer volume of data and the possibility to compare so much through juxtaposition and color variation was deemed by analysts to be “*really clever*.” The appliance based sorting in *Ownership Groups* was seen as both novel and useful: “*The 5 way sorting ... by category, load, subclass is not something we've seen before.*” Initial reactions to the animated transitions in *Demand Horizons* when shifting from weekday to weekend highlighted the novelty of this feature and the sorting of the appliances by their contribution to the peaks was seen as: “*really interesting – you just could not get that out of numbers.*” The data sculpting feature also received positive feedback from analysts suggesting novelty and surprise (see section 6).

Interviews conducted with the data modelers revealed that they also regarded the designs to be novel: “*they give me the opportunity to analyze the data in a different way.*” The designs also enabled the modelers to see surprising structure in their outputs: “*I didn't expect to see these patterns*” and “*I wouldn't be able to spot the problem before I saw this graph.*” The modelers' view on the trial data changed completely upon seeing *Smart Home HeatLines*: “*before I thought the trial data could not be used due to errors and outliers. The visualization showed me that you can use this data and detect different patterns and user behavior.*” There were also clear opportunities identified for data visualization within the energy data modeling domain: “*it has got great potential ... to spot problems, abnormalities, see the patterns, come up with new ideas, new theories, new models.*”

5.3 Process Validity and Effect

The analysts felt engaged in the process, that they had contributed and that they had learned through doing so. They were pleased with the responses to their suggestions: “*you actually listened to our feedback, helped us shape that feedback and then delivered.*” The process of developing the prototypes was deemed to be educational and stimulating helping the analysts understand the possibilities that data visualization can offer and the value of considered visual design: “*I realize that actually this has got many potential applications and many many uses,*” “*the data is a crucial thing and the visualization of that data is almost as important to move ... from information to insight.*”

6 CASE STUDY: DATA SCULPTING

One example of novelty, as perceived by the energy analysts, relates to the ability in the *Demand Horizons* prototype to engage in *data sculpting*. Documenting the lineage of the idea through our development process draws attention to the creative processes and enables us to reflect on the impact of the creativity methods we used.

6.1 Requirements Workshop

It was evident that the potential impact of successfully implementing the *Plan of Action* (see section 3.1.4), that arose out of visualization awareness with analogical reasoning, would be significant in economic

and environmental terms: power stations are costly on both counts. The importance of the power station as a unit of production was also very clear: they are used to accommodate peaks in energy consumption, difficult to switch on and off and expensive to maintain – hence the significance of reducing peaks below the threshold at which a particular plant is needed. We thus took the *Plan of Action* to the *Design Concepts Workshop* as one of our key inputs as we had been informed that: “*the better stage 1 is, the better stage 2 and 3 will be.*” A designer explains how this inspired the development of *Demand Horizons*.

The Designer's Story – Initial Development.

I chose to design to “How can we use visualization to better understand how different appliances contribute to the peaks in energy demand throughout the day?” The objective was to design paper prototypes to meet this requirement without consideration of data or development constraints. Having some experience of developing data visualization techniques and systems, I was keen to make a contribution that fitted technique to requirement in a creative way. Knowing that many appliances might have to be shown concurrently, I was looking for a visual technique that was graphically compact, but visually distinctive. Horizon charts [20] seemed particularly appropriate as energy production jumps between discrete quanta when power stations are fired up or shut down in line with demand. This had a natural fit with the discrete ‘horizons’ of the chart. Thus the initial prototype comprised a set of horizon charts – one per appliance – and a single summed horizon chart representing total consumption. Each discrete band might represent the consumption necessary to cause a power station to be brought online (see Fig. 6).

The modeled data populating the horizon application were somewhat approximate and subject to change as the consumption model changed. This uncertainty informed the smoothed line design of the horizons as well as the smooth transitions implemented when moving between weekend and weekday consumption models (see Fig. 1).

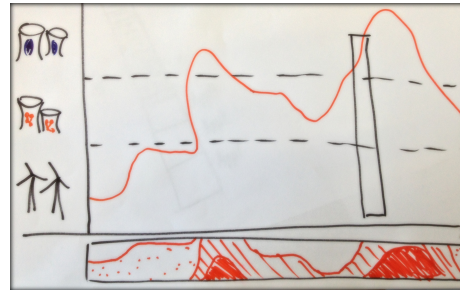


Fig. 6. Design Sketch: The *Demand Horizons* view of the *Plan of Action*.

6.2 A Creative Feedback Workshop

Initial reactions to the prototype at the *Feedback Workshop* were very positive: analysts liked the ability to play with the representation and see things change – a novelty to them in terms of their use of their data: “*there are so many touch points and ways I can move around that data – it gives you a Wow! factor*” and “*I think this is very powerful.*” Switching quickly between standard and horizon graphs helped explain the horizons and we were soon in a position where discussion about the data flowed with theories and requirements explored enthusiastically: “*if this data was live I'd like to be able to look at specific days – i.e. load shifting for tumble dryer could relate to specific days such as [when we have] rain.*” The ability to switch between weekday and weekend consumption was positively received and emphasized the

fluidity of the interface and ability to change the data seamlessly and quickly to suit particular lines of enquiry. This emphasis on fluidity and flexibility seemed to inspire some creative thinking about using the data that gave rise to interesting ideas and subsequent requirements in terms of managing energy consumption – “cooker goes off, dishwasher comes on! Can we shift the dishwasher?” – and important discussion around the timings of usage of washing machines (mainly in the morning) and driers (main usage in the evening): “[could consumers] use the washer, they leave it and then they dry it when they come home?” The significance here is that if consumers are prepared to wait to use energy consuming devices there is scope for offsetting usage to reduce the evening peak – perhaps below a power station horizon.

This exploration of patterns in the modeled data gave rise to further creative thinking about using the interface to model changes in consumption – through changes in behavior and more energy efficient devices: “you can’t shift lighting time ... but we can remove a percentage by changing the bulbs” and “[what if we] switched everyone to a more efficient fridge freezer for example?” The aim of moving the dark peak below the upper horizon was implicit in the vigorous discussion. The design appeared to have been revealing and instructive in focusing activity on the need to reduce consumption below the levels emphasized by our horizons – much in the way anticipated by our designer, and clearly in line with the *Plan of Action*.

In turn these ideas rapidly gave rise to discussion about the interface and how we might interact with data to explore these theories: “could we drag and drop and move something from that time to another time - to imagine [model] time shifting?” We began to explore these ideas collectively: of reducing the consumption profiles of particular devices by a proportion and of moving consumption of particular devices from one time to another to remove the top horizon. Animated discussion ensued in front of the projected images with ideas being developed rapidly about how to select and represent times, percentages and shifts. This was intensive, creative design work inspired directly by data and analytic need, the latter being identified directly prior to the design ideas discussion through our prototype interface. The analysts were excited by their increased understanding and interpretation of the data, design possibilities and new ways of interacting with the models to address their objectives. This was evident in ensuing discussions about deployment and the immediate request for screen dumps to be used in an imminent internal meeting. Our focus here was very definitely on step 2 of the *Plan of Action* – *displacement* – as the data prototype had addressed much of step 1 – *discovery*. The ideas captured during this highly creative discussion at the *Feedback Workshop* were particularly useful as they were stimulated by both interface and the data analysis it enabled, in the context of an identified objective. They were communicated to our developers for the second development iteration.

The Designer’s Story – Enhanced Functionality

The requirement to allow ‘what if?’ remodeling of consumption patterns was clearly expressed, leading to the need to be able to edit the data shown in the horizon charts. Rather than separate the editing from the data exploration tasks, I combined the two processes under the metaphor of ‘data sculpting’, enabling analysts to interactively select time periods and then vary consumption levels for particular appliances with immediate graphical feedback. This idea arose in part from previous work I had seen and developed for ‘sculpting’ terrain models where interactive graphical tools are used to raise and lower parts of a gridded elevation model [39, 40]. It also follows the design pattern of ‘data as interface’ that I had found successful previously [10]. The metaphor was reinforced and partly inspired by the use of a clay colored color scheme and the smooth curves used in the charts that make the graphs look as though they are mouldable.

6.3 Evaluation Workshop

The *data sculpting* feature sparked a vibrant discussion at the *Evaluation Workshop* with plenty of ideas of possible uses. It seems to be a technique with scope for helping explain the concept of demand shifting and reduction and to explore its possibilities: “I am more confident

that internally I could use something like this to demonstrate that it [flexible demand] will work.” Known aspirations for switching cold appliances off and on were discussed, with the interface encouraging new thinking: “the fantastic thing about grow is you can grow before hand as well so you can super cool fridges or freezers.” The feature was deemed “a very useful dynamic tool” that could pave the way for a new data storage strategy to ensure that data is of sufficient resolution to allow for this kind of visualization.

The modelers also liked the idea of *data sculpting* and had not considered using visualization in this way: “this is really good. It represents what we have tried to do with the optimization tool but when I produce a model or amend it we need to re-run it. This does it instantly!” The modelers were positive when asked whether *data sculpting* would be useful to help with building and editing the optimization algorithm itself: “yes, if I had something similar to that I would definitely use that.” New ideas were also created such as relating the horizons to energy cost thresholds: “if the cost exceeds the thresholds you would have a penalty. You could visualize it and see it.”

7 REFLECTION

The evaluation and case study reported above demonstrate some success in terms of our applied designs. Approaches such as *data sculpting* in *Demand Horizons* and the comparison through color variation and alignment used in *Consumption Signatures* and the multi-scale interactive analysis through *Smart Home HeatLines* demonstrate some novelty, seem useful in this context and may be applicable in other domains and scenarios. In this section, we share the reflections of experienced designers on the extent to which the process we have undertaken can be seen as *creative*, and consider the impacts this may have had on designers and other stakeholders.

In an applied client-based project such as this, evaluating the impact of the methods used by means of a controlled study is not feasible. Our approach to gaining some initial insights on the impact of our creative methods on the process of visualization design has therefore been to reflect, as designers, on our experience in this project, in order to compare it with the numerous other projects in which we have been involved over the years. Without a control we are unable to prove that adding the creativity methods at the outset of the project had any specific impacts on the process as a whole: good visualization design projects almost always involve creativity and novelty and we actively emphasized and valued these characteristics here. However, we did feel that the creativity methods opened up particular opportunities for creative thinking. They established the true breadth of a situation in which requirements are open with familiar reference points. They took participants out of their comfort zones and enhanced the ‘away day effect’ of shared purpose. The explicitly creative activities helped visualization designers and domain experts communicate, share experiences, establish trust and work as a team. We experienced creative thinking about using data as well as about design and the creative thinking may indeed have helped us “push domain experts to discuss problems, not solutions” [47]. Based on our experience of past projects, we identify the elements where we feel the use of deliberate creativity methods had the greatest impact in Fig. 2 and discuss these further below.

Some of the simplest creativity methods seemed surprisingly effective. The animal introductions required some audacity on the part of our facilitator, but this was handled with aplomb. Developing analogies and revealing some personal information in a controlled and safe manner required openness on behalf of all participants. It seemed useful preparation for future exercises in initially putting all participants on an equal footing, establishing trust and involving surprise – suggesting that anything was possible from the outset. The excursion worked well as a preparation exercise to get participants in the frame of mind for the next activity and remind them that lunch was an opportunity to think and communicate. Everyone understood, brought something interesting back and had time to make a contribution.

Our impression following the visualization awareness activity was that use of analogy was very evident. Participants applied many of the ideas shown in visualizations from other domains creatively and effectively to their own area of interest. This activity spurred on a long and

interesting conversation about what was possible with the data to hand and might be achievable given the visualization examples presented. It seemed that these ideas generated after the visualization demos were stimulated by the morning's activities. We regarded them to be more numerous and creative than is the norm in these sessions and the outputs – such as mind maps developed during the awareness activity – were sophisticated. The storyboards produced in the activity that followed were not as useful as we had hoped. This may have been due to a lack of energy or the fact that previous discussions meant that we were overrunning – partly because graphical summaries were already being produced as participants took the initiative to generate mind maps in response to the analogical reasoning activity. Sketches or stories that are more data focussed may be more useful in our domain and we are likely to encourage the mind-mapping as a visualization storyboard during analogically focussed awareness activity in the future.

The novel ideas established at the subsequent *Feedback Workshop* are not easy to attribute directly to the initial use of creativity methods, but were rare in our experience of user-centered visualization design in terms of their quality, relevance and originality. The expressions of novelty and surprise (see section 5.2) were particularly embedded in organizational context, including evidence of insights, and realizations of new capacity and scope for the group. Possible changes in the way that the organization stores and uses data were suggested. Our sense was of a strong link and our activity felt focussed with participants particularly engaged and able to make excellent and sometimes unexpected suggestions for design possibilities throughout the process. We claim above that creativity may have persisted throughout the one-day *Requirements Workshop*. We also suggest that the early use of creativity methods may have had longer lasting effect through our study. Equally, being explicit about our desire and efforts to be creative may have been beneficial – a positive example (in design terms if not in experimental terms) of the experimenter effect in an *in vivo* situation where controls are not feasible. The *Designer's Story* (see section 6.1) offers some evidence to support this suggestion.

In terms of process, the analysts felt that they had made beneficial contributions and been able to communicate effectively with the design team. They reported benefits in terms of both understanding the data and visualization possibilities (sections 3.1.6 and 5.3). We felt that levels of engagement and learning were high and would associate this with the persistent sense of creativity that we are reporting. We acknowledge that this sense of contribution and ownership may have an effect on the evaluation – a positive bias being highly likely. However, it may also have an effect on uptake, which could be evaluated through a longitudinal study post implementation [48].

Our designs were not wholly successful in terms of analyst reactions however. *Ownership Groups* quickly revealed that the modeled data did not capture the kinds of relationships between users and appliances that we had hoped to explore. The lack of a realistic pattern emerging meant that analysts were less engaged with this application than the others, reinforcing established findings [27]. Reflecting back, it seems that we may have been collectively over-optimistic in anticipating that we could either find or imagine patterns where our data did not support them (see the data description in section 2). Perhaps the creative nature of our *Design Workshop* resulted in some inefficiency and inappropriate design. Perhaps explicitly creative visualization design processes may produce more 'misses' than standard approaches and thus be particularly costly. Perhaps – but benefits may also be associated with this cost. We captured plenty of suggestions that our prototypes were relevant beyond the original use cases and target group (see section 5); with various ideas for *Smart Home Heat-Lines* and *Demand Horizons* being used in other organizational and customer facing contexts. Additionally our designs were deemed useful by the modelers, who used them to develop insights and expressed interest in building aspects of the prototypes into their workflows (see section 6.3). We are unable to establish whether this is due to the open requirements, unknown data and design to focus points rather than formal task analysis (all used in previous design studies), or the parallel design or the creative approaches used in this case. Further work is needed to explore these various possibilities and any effects.

8 CONCLUSION

Our experience of using deliberate creativity techniques in the visualization design process has been very positive. We present reactions from the domain experts – energy analysts and data modelers – and reflect on our own experiences to support this view. We describe a series of candidate designs for energy visualization that have been developed through intensive user-centered collaboration. They have been enthusiastically received in most cases in light of initial requirements and expectations and have resulted in insights about data, new knowledge about analytical and visualization possibilities and potential behavior change in individuals and within organizations. They may be more widely useful as energy visualization becomes more widespread. Our evaluation supports the conclusion that they constitute a successful exploration of possibilities for analytical Smart Home data visualization.

Energy analysts and modelers found the designs novel and useful. Designers also developed methods they deemed novel in collaboration with and response to analysts. We claim, through reflection informed by our experience of what has been a lengthy and intense process, that *the explicit use of creativity methods is likely to have contributed to the development of novel and effective solutions that are well aligned with established need*. This is particularly significant in a situation where requirements are open and data largely unknown. We cannot trace back through the hundreds of prioritized requirements and captured reactions, the hours of discussion and the piles of sketches to establish a direct causal link between the creativity sessions and our designs – we don't think this is how it works. Visualization design is much more holistic, taking ideas from all sorts of influences often in parallel – just as good visual thinking uses multiple stimuli concurrently to generate ideas and make decisions. Indeed, we suspect that *the very fact that we were explicit from the outset about creativity being a focus in the project may well have made us more creative in our approaches*. The *Designer's Story* (section 6.1) suggests that this may well be the case.

We conclude that *the deliberate use of techniques to enhance creativity early in the visualization design process can contribute to success in terms of process and outcomes*. In our experience this proved highly likely to be the case in: establishing a creative working environment; developing requirements; pushing designers and developers to novel solutions; and building a sense of trust, common purpose and ultimately achievement in a diverse team. Furthermore we suggest that *using creativity techniques early in the visualization design process may have longer term positive effects on creativity and satisfaction that persist throughout a design process and perhaps beyond*.

In applied design projects domain experts' time is limited and valuable. We find real benefit in encouraging them to be as creative as possible early in the process as our experience suggests that creative methods challenge mental and social barriers, can enthuse and energize participants and engage them in design. Carefully facilitated, visualization focussed, use of *wishful thinking*, *constraint removal*, *excursion*, *analogical reasoning* and *reflection* may be straightforward 'discount' methods that contribute to buy-in, satisfaction and the efficient use of participants' time. We see room for using these creativity techniques and others, such as creativity through random combination [37], at various stages through the design process to explore their effects. Indeed, we plan to use creativity techniques in future projects as they seem to provide a low cost means of establishing a beneficial creative climate. We call on others to do the same. Perhaps documenting and reflecting upon the creative aspects of the design and indeed analytical processes in a series of projects will be the best way to share and assess experiences. We may then begin to understand more about the specific effects of creativity on user-centered visualization design.

ACKNOWLEDGMENTS

This work was undertaken by City University London and the IMDEA Energy Institute, Madrid through E.ON AG International Research Initiative (IRI) 2012. Thanks to Amanda Brown, Jorn Gruber, Soroush Jahromizadeh, Milan Prodanovic, Veselin Rakocovic and the Forward Thinking Technologies Team at E.ON UK for contributions to the study, Nabihah Ahmed for producing the video and Miriah Meyer and four reviewers for useful suggestions that have improved the paper.

REFERENCES

- [1] N. Bonnardel. Creativity in design activities: The role of analogies in a constrained cognitive environment. In *Proceedings of Creativity & Cognition 3*, pages 158–165. ACM, 1999.
- [2] K. Brennan. *A Guide to the Business Analysis Body of Knowledge*. International Institute of Business Analysis, 2nd edition, 2009.
- [3] B. Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design*. San Francisco: Morgan Kaufmann, 2007.
- [4] C. Clastres. Smart grids: Another step towards competition, energy security and climate change objectives. *Energy Policy*, 39:5399–5408, 2011.
- [5] M. Cohn. Techniques for estimating. In *Agile Estimating and Planning*, pages 49–60. Addison-Wesley, Boston, 2005.
- [6] N. Cross. Creative cognition in design: Processes of exceptional designers. In *Proceedings of Creativity & Cognition 4*, pages 14–19. ACM, 2002.
- [7] Cruz, P. Empires Decline: Revisited - (<http://bit.ly/10qlaea>), 2010.
- [8] D. Dean, J. Hender, T. Rodgers, and E. Santanen. Identifying quality, novel, and creative ideas: Constructs and scales for idea evaluation. *Journal of the Assoc. for Information Systems*, 7(10):649–699, Oct. 2006.
- [9] S. P. Dow, A. Glasco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction*, 17(4):1–24, Dec. 2010.
- [10] J. Dykes, J. Wood, and A. Slingsby. Rethinking map legends with visualization. *IEEE TVCG*, 16(6):890–899, 2010.
- [11] G. Ekvall, J. Arvonen, and I. Waldenström-Lindblad. *Creative Organizational Climate: Construction and Validation of a Measuring Instrument*. Swedish Council for Management and Organizational Behaviour, 1983.
- [12] K. Ellegård and J. Palm. Visualizing energy consumption activities as a tool for making everyday life more sustainable. *Applied Energy*, 88:1920–1926, 2011.
- [13] A. Faruqi, D. Harris, and R. Hledik. Unlocking the 53 billion euro savings from smart meters in the EU. *Energy Policy*, 38:6222–6231, 2010.
- [14] S. Firth, K. Lomas, A. Wright, and R. Wall. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings*, 40(5):926–936, Jan. 2008.
- [15] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.
- [16] W. J. Gordon. *J.(1961) Syntetics: The Development of Creative Capacity*. New York: Harper & Row, 1960.
- [17] J. Gruber and M. Prodanovic. Residential energy load profile generation using a probabilistic approach. In *6th European Symposium on Computer Modeling and Simulation*, pages 317–322, Valetta, Malta, Nov. 2012.
- [18] T. Hargreaves, M. Nye, and J. Burgess. Making Energy Visible: A Qualitative Field Study of How Householders Interact with Feedback from Smart Energy Monitors. *Energy Policy*, 38(10):6111–6119, Oct. 2010.
- [19] M. Harrower and C. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal*, 40(1):27–37, 2003.
- [20] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of Human Factors in Computer Systems*, pages 1303–1312. ACM, 2009.
- [21] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE TVCG*, 13(6):1240–1247, Nov. 2007.
- [22] L. Hohmann. *Innovation Games: Creating Breakthrough Products Through Collaborative Play*. Boston: Addison-Wesley, 2007.
- [23] S. G. Isaksen, K. J. Lauer, and G. Ekvall. Situational outlook questionnaire: A measure of the climate for creativity and change. *Psychological Reports*, 85(2):665–674, 1999.
- [24] S. Jones, P. Lynch, N. Maiden, and S. Lindstaedt. Use and influence of creative ideas and requirements for a work-integrated learning system. In *16th IEEE International Conference on Requirements Engineering*, pages 289–294. IEEE, 2008.
- [25] L. Koh, A. Slingsby, J. Dykes, and T. Kam. Developing and applying a user-centered model for the design and implementation of information visualization tools. In *15th International Conference on Information Visualisation*, pages 90–95, London, 2011. IEEE.
- [26] D. Lloyd. *Evaluating Human-Centered Approaches for Geovisualization*. PhD thesis, City University London, 2009.
- [27] D. Lloyd and J. Dykes. Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. *IEEE TVCG*, 17(12):2498–2507, 2011.
- [28] M. Maher and D. Fisher. Using AI to evaluate creative designs. In *2nd International Conference on Design Creativity*, Glasgow, UK, Sept. 2012.
- [29] N. Maiden, A. Gizikis, and S. Robertson. Provoking creativity: Imagine what your requirements could be like. *IEEE Software*, 21(5):68–75, 2004.
- [30] N. Maiden, S. Manning, S. Robertson, and J. Greenwood. Integrating creativity workshops into structured requirements processes. In *Proceedings of 5th Conference on DIS*, pages 113–122. ACM, 2004.
- [31] N. Maiden, C. Neube, and S. Robertson. Can requirements be creative? experiences with an enhanced air space management system. In *29th IEEE International Conference on ICSE*, pages 632–641, 2007.
- [32] N. Maiden and S. Robertson. Developing use cases and scenarios in the requirements process. In *Proceedings of 27th International Conference on Software Engineering*, pages 561–570. ACM, 2005.
- [33] E. McFadzean. The creativity continuum: Towards a classification of creative problem solving techniques. *Creativity and Innovation Management*, 7(3):131–139, 1998.
- [34] M. Michalko. *Thinkertoys: A Handbook of Creative-Thinking Techniques*. California: Ten Speed Press, Dec. 2010.
- [35] M. J. Muller and S. Kuhn. Participatory design. *Communications of the ACM*, 36(6):24–28, 1993.
- [36] A. F. Osborn. *Applied Imagination, Principles and Procedures of Creative Thinking*. New York: Scribner, 1953.
- [37] A. F. Osborn. *Applied imagination: Principles and Procedures of Creative Problem-Solving*. New York: Scribner, Rev. ed edition, 1957.
- [38] L. Pennell and N. Maiden. Creating requirements—techniques and experiences in the policing domain. In *Proceedings of REFS 2003 Workshop*, 2003.
- [39] Pixologic Inc. Pixologic :: Sculptis - (<http://bit.ly/lymthei>), 2013.
- [40] PlanetSide Software. Terragen 2 - (<http://bit.ly/10n2b5o>), undated.
- [41] R. Radburn, J. Dykes, and J. Wood. vizLib: Using the seven stages of visualization to explore population trends and processes in local authority research. *Proceedings of GIS Research UK*, pages 409–416, 2010.
- [42] J. Robertson. Eureka! Why analysts should invent requirements. *IEEE Software*, 19(4):20–22, 2002.
- [43] J. Rodgers and L. Bartram. Exploring Ambient and Artistic Visualization for Residential Energy Use Feedback. *IEEE TVCG*, 17(12):2489–2497, Dec. 2011.
- [44] S. Rusitschka, K. Eger, and C. Gerdes. Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain. In *1st IEEE International Conference on Smart Grid Communications*, pages 483–488, 2010.
- [45] E. B.-N. Sanders. Information, inspiration and co-creation. In *Proceedings of 6th International Conference of European Academy of Design*, 2005.
- [46] K. Schmid. A study on creativity in requirements engineering. *Softwaretechnik-Trends*, 26(1):20–21, 2006.
- [47] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG*, 18(12):2431–2440, Dec. 2012.
- [48] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of BELIV - Beyond time and errors: novel evaluation methods for Information Visualization*, pages 1–7. ACM, 2006.
- [49] A. Slingsby and J. Dykes. Experiences in involving analysts in visualization design. In *Proceedings of BELIV - Beyond time and errors: novel evaluation methods for Information Visualization*, page 1. ACM, 2012.
- [50] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.
- [51] J. W. Tukey. *Exploratory data analysis*. Reading, MA, 231, 1977.
- [52] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of 1999 IEEE Symposium on InfoVis*, pages 4–9, 1999.
- [53] A. Warr and E. O'Neill. Understanding design as a social creative process. In *Proceedings of Creativity & Cognition 5*, pages 118–127. ACM, 2005.
- [54] J. Wood, A. Slingsby, and J. Dykes. Using treemaps for variable selection in spatio-temporal visualization. *Information Visualization*, 7(3):4, 2008.
- [55] Wood, J. Experiments in bicycle flow animation - (<http://bit.ly/10f2jie>), 2012.
- [56] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evans. *R66141 Final Report Issue 4: Household Electricity Survey: A Study of Domestic Electrical Product Usage*. May 2012.

A.2 Short Paper: IEEE VIS 2014, Paris, FR

*Preprint paper for IEEE Conference on Visual Analytics Science and Technology,
09-11-2014 - 14-11-2014, Paris, France.
Available here: <http://openaccess.city.ac.uk/3876/>*

Visualizing the Effects of Scale and Geography in Multivariate Comparison

Sarah Goodwin*

Jason Dykes†

Aidan Slingsby‡

giCentre, City University London

ABSTRACT

Our research investigates the sensitivities and complexities of visualizing multivariate data over multiple scales with the consideration of local geography. We investigate this in the context of creating geodemographic classifications, where multivariate comparison for the variable selection process is an important, yet time-consuming and intensive process. We propose a visual interactive approach which allows skewed variables and those with strong correlations to be quickly identified and investigated and the geography of multi-scale correlation to be explored. Our objective is to present comprehensive documentation of the parameter space prior to the development of the visualization tools to help explore it.

Index Terms: D.2.2 [Software Engineering]: Design Tools and Techniques; I.5.2 [Pattern Recognition]: Design Methodology—Feature Evaluation and Selection

1 INTRODUCTION

The comparison of geographically varying phenomena is both position and scale dependent. We investigate this in the context of creating and visualizing geodemographic classifications. Geodemographics group geographical areas by similar population characteristics and are used by academics, governments and professionals to identify typical population or customer characteristics [5].

The selection of variables through comparison is an important part of building the classifier and variables should be independent, of near-normal distribution and have little or no correlation to one another [5]. The variable selection (known in clustering as ‘feature selection’ [6]) is a time consuming and intensive process [5, 13], which may be subjective to user interpretation. We propose a visual interactive approach to aid the process, allowing skewed and strongly correlating variables to be quickly identified and investigated and the geography of multi-scale correlation to be explored.

Scale and geography are of particular importance in our proposal as knowledge of local variations may influence variable selection and classifications can be created at multiple scales with each likely to produce very different outcomes. There is limited research in the area of spatially weighted geodemographics [1] or varying geodemographic scales. Our research investigates the sensitivities and complexities of visualizing multiple data variables over multiple scales with the consideration of local geography.

2 DATA SOURCES

This research follows previous work on investigating domain specific geodemographic visualization and creation in the context of energy consumption [3]. We use small-area summary statistics from the 2011 UK Census [9], based on the open geodemographic

*e-mail: Sarah.Goodwin.1@city.ac.uk

†e-mail: J.Dykes@city.ac.uk

‡e-mail: Aidan.Slingsby.1@city.ac.uk

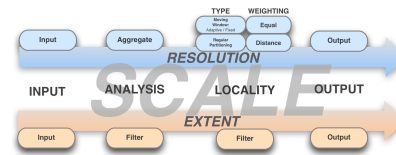


Figure 1: Four stages of the process: Input, Analysis, Locality and Output, each with two dimensions of Scale: Resolution and Extent

methodology [13], combined with energy consumption data for gas and electricity from the ‘Department of Energy and Climate Change’ (DECC) [2].

3 SCALE

Figure 1 identifies four stages of the variable selection process in which scale can be varied: *Input*, *Analysis*, *Locality* and *Output*. Adjusting the scale particularly at the two central stages allows the associated sensitivities to be explored. At each stage there are two dimensions: *Scale Resolution* and *Scale Extent* [7, 12], which are defined as:

Scale Resolution (SR) - the level of aggregation used to make comparisons. When data is aggregated the nature of the summaries used to describe areas at each scale, and relationships between them, can vary. Aggregation of data can remove outliers and is associated with the modifiable areal unit problem (MAUP) [11]. The use of visualization to illustrate how different variables react to changes in resolution may help to identify the optimal resolution for analysis as well as illustrate the effects of MAUP.

Scale Extent (SE) - the geographical extent of the data; for example selecting the whole of the dataset or a subset (a geographic filter) of the data can lead to entirely different results.

The four stages introduced above, can be defined as follows:

Input - resolution (IR) and extent (IE) refers to the smallest areal unit and full extent of the ‘raw data’. For our data sources this is Output Area [10] for the Census variables and Lower Super Output Area [10] for DECC. Both sources have an IE that covers England and Wales.

Analysis - resolution (AR) and extent (AE) refers to the scale for the chosen analysis. The IR may be aggregated to a larger areal unit for example Local Authority region (AR) and/or the IE can be filtered to a specific geographical area of interest (AE), such as Wales or Greater London.

Locality - resolution (LR) and extent (LE) allows for the calculation of summary statistics at varying local as well as global scales. Such local summary statistics can be calculated in various ways as indicated by *Type* in Fig. 1. These include using a *Moving Window* technique with a *Fixed* (number of areas) or *Adaptive* (using a distance measurement) kernel or by using *Regular Partitioning*, where a grid (of a certain distance) is overlaid on to the data (size > AR).

Weighting refers to whether the areal units within the moving window or partition are given equal or distance weighting to the cal-

Table 1: Table identifies the ability to make comparisons when visualizing multiple Scale Resolutions (SR) and Extents (SE) with increasing numbers of variables (V) and local summaries (L)

	Distribution	Correlation	
	V=1	V=small	V=large
L=1	SR: Many SE: Many	SR: Some SE: Limited	SR: Limited SE: None
L=small	SR: Many SE: Many	SR: Some/Limited SE: Some/Limited	SR: Limited/None SE: Limited/None
L=large	SR: Many SE: Many	SR: Limited/None SE: Limited/None	SR: None SE: None

culuation of the local statistic. This framework is based on the principles of Geographically Weighted Modelling [4]. LE is changed from AE only if locally weighted statistics are needed in a subset of the analysis, for example to investigate locally weighted statistics in London compared to elsewhere.

Output - resolution (OR) and extent (OE) refers to the dimensions of the data once it has been through the previous stages and is ready for spatial aggregation to a lower resolution. OR = AR unless *Partitioning* has been chosen in *Locality* then OR will take the size of the partition. OE = AE, unless LE has been utilized.

4 VISUAL COMPARISON

Through the utilization of *Locality* we can calculate local as well as global summary statistics for each variable and with this the complexity of the visualization options increase. The visual representation of such a complex set of scales can be simplified by considering scale in three broad and loosely delimited bands: global (as used in cases where local variations are not considered), macro and micro. Where L = 1 for Global, L = small (but >1) for macro and L = large for micro. The point at which macro becomes micro depends upon the number of variables being shown (V), the number of data points in the comparison, the visualization represented and the users' experience and display possibilities. The ability to make comparisons when exploring the parameter space reduces with increased V and L, as shown in Table 1. This ability to explore the data must be reflected in an adaption of the visual representation at these thresholds. Possibilities for visually encoding these data are multifarious. Given the need to compare skewness of variables and strong correlations both globally and locally we propose two types of visual representation: Statistical and Spatial as shown in Table 2.

4.1 Statistical and Spatial Views

As shown in Table 2 when V and L are large presenting a detailed comparison visually becomes difficult and here we rely on color encoding of the correlation coefficient (or other descriptive statistics in the case of V=1) for a space efficient representation. Matrices in which cells represent pairs of variables can be useful in the layout - whether this is through multiple scatterplots [8], maps showing the geographies of correlation of all pairs of variables or a color encoded grid cell showing the global level of association between each pair. Asymmetrical matrices have been identified as a possible way to compare two differing datasets: for example before and after a data transformation.

5 CONCLUSION

Having established the need for visual representation to support the sensitive and time-consuming issue of variable selection we have produced a framework for considering and visualizing the multiple dimensions of scale and the effects of geography in this process. An interactive application through which these effects can be explored through this framework is in development with novel candidate designs established. Our poster uses the framework to present

Table 2: Table identifying Statistical (top) and Spatial (bottom) visualization possibilities when considering a balance between number of variables (V) and number of local summaries (L). Characteristics of display, user, task and data will be influential in establishing appropriate methods in specific cases

	Distribution	Correlation	
	V=1	V=small	V=large
L=1	Histogram with dot plot	Matrix of Scatterplots	Color encoding
	Choropleth Map (Cartogram or Treemap)	Series of Choropleth Maps	Color encoding
L=Small	Boxplots or Histograms	Matrix of Scatterplots (showing L)	Color encoding
	Choropleth Map	Matrix of Correlation Maps	Matrix of Correlation Maps
L=Large	Color encoding	Color encoding	Color encoding
	Choropleth Map	Matrix of Correlation Maps	Color encoding

these designs graphically, describe the prototype through which the framework is explored and offer reflection and a discussion of opportunities for improvement and future work.

ACKNOWLEDGEMENTS

This PhD research is funded by a Vice Chancellor's Scholarship from City University London and undertaken through collaboration with the g2Lab, Hafencity University.

REFERENCES

- [1] M. Adnan, A. Singleton, and P. Longley. Spatially weighted geodemographics. In *GIS Research UK 21st Annual Conference*, Liverpool University, Apr. 2013.
- [2] DECC. Sub-National Electricity and Gas Consumption Statistics - <http://bit.ly/1bCqsb9>, 2013.
- [3] S. Goodwin and J. Dykes. Visualising variations in household energy consumption. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 217–218. IEEE, 2012.
- [4] P. Harris, C. Brunsdon, and M. Charlton. Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736, 2011.
- [5] R. Harris, P. Sleight, and W. R. *Geodemographics: GIS and Neighbourhood Targeting*. Wiley-Blackwell, 2005.
- [6] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [7] N. Lamand and D. A. Quattrochi. On the issues of scale, resolution and fractal analysis in the mapping sciences. *The Professional Geographer*, 44(1):88–98, 1992.
- [8] M. Monmonier. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21(1):81–84, 1989.
- [9] ONS. Census Data: <http://bit.ly/onsCen11>, 2011.
- [10] ONS. Census Geographies: <http://bit.ly/cenGeog>, 2011.
- [11] S. Openshaw and P. Taylor. *The modifiable unit areal problem*. Norwich:Geobooks, 1984.
- [12] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes. Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Transactions on Visualization and Computer Graphics*, Dec 2014.
- [13] D. Vickers and P. Rees. Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403, 2007.

A.2.1 Poster: IEEE VIS 2014, Paris, FR

Visualizing the effects of Scale and Geography in Multivariate Comparison

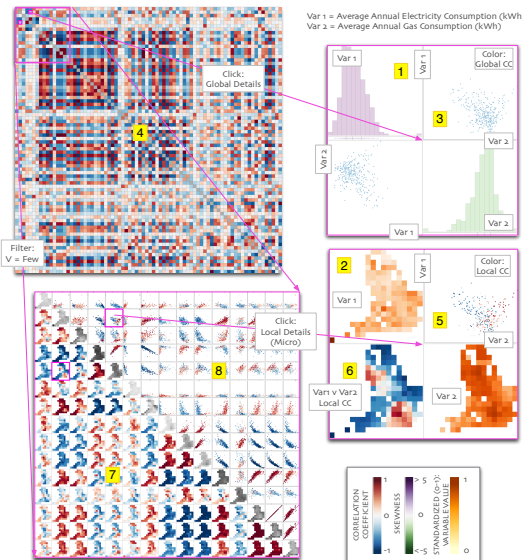
@sgeoviz
Sarah.Goodwin.1@city.ac.uk

Sarah Goodwin, Jason Dykes and Aidan Slingsby - giCentre, City University London, UK

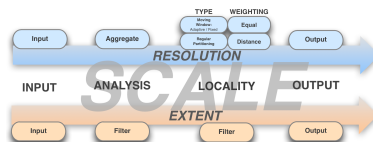
The comparison of geographically varying phenomena is both position and scale dependent. We propose a framework to compare multivariate data across multiple scales whilst allowing local geographical variations and the effects of their parameters to be explored. The statistical (top row) and geographical (bottom row) visualization possibilities adapt as the number of variables (V) and/or the number of local summary statistics (L) increase. An interactive visualization prototype has been built to demonstrate the framework, with aspects (1 - 8) highlighted. Matrices of scatterplots [1], correlation maps or color encoded statistical values (i.e correlation coefficient (CC) and skewness) are utilized to allow as many V as possible to be shown in one view:

Visualizing the FRAMEWORK:

	DISTRIBUTION		CORRELATION			
	V = 1	V = 2	V = FEW	V = MANY		
GLOBAL L = 1	1 Histogram or Dot/Box Plot Map (Choropleth) of Raw Values	2	3 Scatterplot Pair of Maps or Difference Map	4 Scatterplot Matrix Series of Maps or Difference Maps	4 Color Encoding Color Encoding	
MACRO-LOCAL L = FEW	Series of Histograms or Dot/Box Plots Map (Choropleth) of L Values	3	Scatterplot colored by L Values Pair of Maps or Correlation Map	Scatterplot Matrix colored by L Values Correlation Map Matrix	4	4
MICRO-LOCAL L = MANY	Dot/Box Plots or Color Encoding Map (Choropleth) of L Values	5	Scatterplot colored by L Values Pair of Maps or Correlation Map	6	7	8



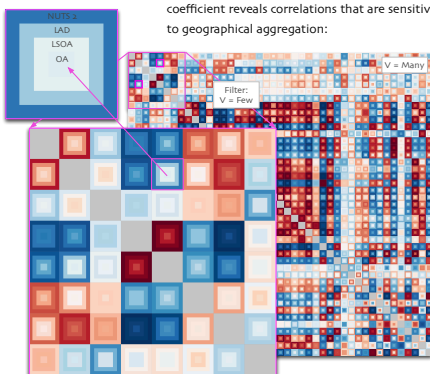
Varying SCALE RESOLUTION:



The framework is investigated in the context of 'variable selection' for energy-based geodemographic classification [2,3]. Four stages (Input, Analysis, Locality and Output) are identified in which Scale Resolution (SR) and Extent (SE) [4,5] can be varied. 78 variables are compared over 4 geographical aggregations (SR) covering a SE of England: NUTS2 European Regions (30), Local Authority Districts (326 LADs), Lower Super Output Areas (32,844 LSOAs) and Output Areas (171,372 OAs). The sensitivities and complexities of varying SR are investigated through visualization:

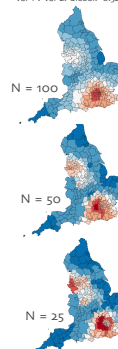
AGGREGATION

Visualizing the color encoded global correlation coefficient reveals correlations that are sensitive to geographical aggregation:



LOCALITY

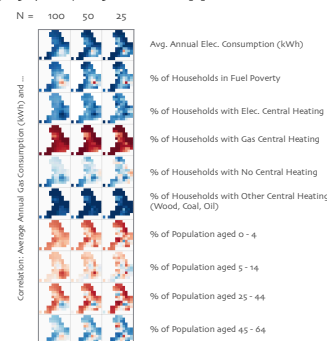
Var 1 v Var 2: Global: -0.32



.... allows for the calculation of local summary statistics (macro or micro). TYPEs are based on geographically weighted statistics [6].

Here, an adaptive moving window is used to calculate local correlation coefficients.

The number (N) of nearest neighbours is adjusted to reveal more detailed local variations when comparing variable pairs (left) or multiple (right):



SUMMARY

This framework and prototype visualization enables the visual comparison of multivariate distributions and correlations across geographical scales and allows for local variations to be explored. The visual representations used in this case-study can be adapted to compare the effects of scale resolution and scale extent that occur when we aggregate and filter by time or attribute as well as geography in our analysis.



CITY UNIVERSITY
LONDON

gicentre.org
Conference Paper: openaccess.city.ac.uk/3876

- [1] M. Monmonier. Geographic Brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 24(1): 81-84, 1989.
- [2] S. Goodwin and J. Dykes. Visualizing Variations in Household Energy Consumption. In *IEEE Conference on VAST 2012*, pages 217-218, Oct 2012.
- [3] R. Harris, P. Sleight, and R. Webber. *Geodemographics: GIS and Neighbourhood Targeting*. Wiley-Blackwell, 2005.
- [4] N. Lam and D. A. Quattrochi. On the Issues of Scale, Resolution and Fractal Analysis in the Mapping Sciences. *The Professional Geographer*, 44(1):88-98, 1992.
- [5] C. Turckay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes. Attribute signatures: Dynamic Visual Summaries for Analyzing Multivariate Geographical Data. *IEEE TVCG*, Dec 2014.
- [6] P. Harris, C. Brunton, and M. Charlton. Geographically Weighted Principal Components Analysis. *Int. Journal of Geographical Information Science*, 25(10): 1717-1736, 2011.

A.3 Short Paper: IEEE VIS 2012, Seattle, USA

*Preprint paper for IEEE Conference on Visual Analytics Science and Technology (VAST),
14 - 19 Oct 2012, Seattle, Washington, US.
Available here: <http://openaccess.city.ac.uk/1294/>*

Visualising Variations in Household Energy Consumption

Sarah Goodwin*

Jason Dykes†

giCentre, City University London, UK

ABSTRACT

There is limited understanding of the relationship between neighbourhoods, demographic characteristics and domestic energy consumption habits. We report upon research that combines datasets relating to household energy use with geodemographics to enable better understanding of UK energy user types. A novel interactive interface is planned to evaluate the performance of specifically created energy-based data classifications. The research aims to help local governments and the energy industry in targeting households and populations for new energy saving schemes and in improving efforts to promote sustainable energy consumption. The new classifications may also stimulate consumption awareness amongst domestic users. This poster reports on initial visual findings and describes the research methodology, data sources and future visualisation requirements.

Index Terms: I.3.8 [Computing Methodologies]: Computer Graphics—Applications; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; H.2.8 [Information Systems]: Database Management—Database Applications;

1 INTRODUCTION

Energy consumption and carbon footprint reduction is of growing interest to individuals, organisations and government. Added pressure of EU carbon and emission targets set for 2020 and 2050 has meant that tackling energy consumption in the UK is of major concern. Reducing domestic energy consumption in particular is challenging due to large variations in household energy use. Patterns and trends in consumption levels in relation to housing, population, lifestyle and behaviours must be better understood in order to implement successful strategies in a movement to achieve efficient, sustainable and low carbon residential living environments.

Over recent years there has been a growing amount of academic and government-led research focused on energy consumption, carbon reduction and potential energy saving schemes; however, there is still limited knowledge of the relationship between consumption and measurable characteristics of the population. It is reported that UK domestic fuel consumption is strongly related to disposable income levels with other highly influential factors being dwelling type, household composition, property tenure and rural/urban location [4]. These findings, along with other research in the field, indicate that energy consumption patterns correlate to socio-economic and geographic characteristics and continued research in this area is needed in order to better understand the complex variations, allow for realistic comparisons amongst neighbours and facilitate better targeting of services and schemes. In this research we use data classification methods, analytical techniques and data visualisation to aid the interpretation and discovery of geographic and demographic variations in UK domestic energy consumption.

* e-mail: sarah.goodwin.1@city.ac.uk

† e-mail: j.dykes@city.ac.uk

2 GEODEMOGRAPHICS

Despite there being a body of relevant research correlating energy consumption with household or population variables, little research directly investigates the classification and evaluation of energy related variables with geodemographics. Geodemographic classification systems are used by geographers, policy makers and market analysts alike to segment the general population and identify trends and patterns based on typical user traits. Geodemographic data products such as Experian's Mosaic [6] or the free and open alternative Output Area Classification (OAC) [10], based entirely on the UK Census output, can greatly enrich consumer databases.

Large multi-variate dataset classification is ideal for residential energy consumption as previous research shows that human populations with similar characteristics and behaviours tend to cluster together. A research study [4] comparing consumption data with the 7 OAC Supergroups reveals variations between groups and clear correlations to household disposable income, property tenure and rural/urban location. We identify similar patterns when combining average electricity consumption data [3] with the 15 Mosaic Groups; Figure 1 reveals that the groups labelled 'Affluent' or 'Rural' display a higher average consumption (darker orange) than the others. The spatial variation of consumption within these groups is of particular interest as it highlights that energy use also varies geographically within these demographic clusters.

The selected datasets, weightings and methodology used for the classification process can all contribute to bias during data clustering. To reduce this bias it is necessary to use variables known to be relevant to the specific domain. As a correlation has been identified when combining geodemographics with energy consumption data, we propose that the geographical clustered of energy consumption data together with relevant variables of household and population characteristics could greatly improve the interpretation of both demographic and geographic variations in household energy usage. Such a energy-specific classification follows a recent call for geodemographics to be brought into the current data and technical era and follow more domain specific, problem centred approaches utilising the advances in visualisation and data exploration techniques [8].

3 SMARTER TECHNOLOGIES AND FEEDBACK

Within the energy industry there is already a substantial amount of household consumption data suitable for building an energy-based classification. The introduction of modern smart meters will; however, increase this data quantity exponentially. Smart meter technology allows for consumption to be recorded at frequent intervals and communicates this information to both consumer and energy supplier allowing for near real-time feedback of energy use. The introduction of smart meter technologies is expected to improve household consumption awareness as well as allow for better regulation of household energy demand. Smart meters form a major component of Smart Grids, which are estimated to reduce annual EU household consumption by 10% and carbon dioxide emissions by 9% [5]. Smart meter data could potentially improve our energy-based classifications for example introducing clusters related to high demand at certain times of the day or days of the week. In order to assess the potential for this granular dataset we are currently working in collaboration with the energy utility company E.ON on a project entitled 'Visualising the Smart Home'.

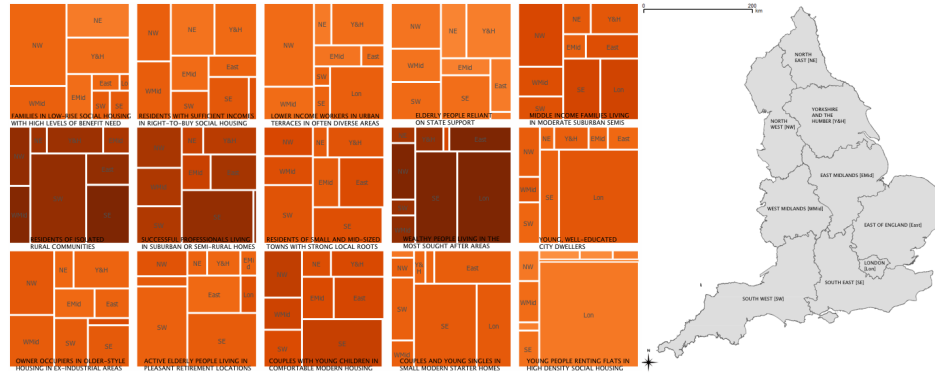


Figure 1: Mosaic Groups mapping English government regions sized by number of electricity meters, coloured by average electricity consumption

Increased pressure to reduce consumption combined with advances in meter technology has seen a growing need and demand for householders to be provided with more transparent and detailed understanding of energy use within their home. A significant amount of research is available evaluating alternative methods of consumer feedback. Relatively recent research identifies that social norms can be used effectively in the case of energy consumption reduction [1]. The research investigates the campaigns by *oPower* in the US, which compare households to a collection of neighbours with similar characteristics, and concludes that a small but continuous and sustained consumption reduction is achieved. These findings suggest that neighbourhood level comparisons based on geodemographic energy classifications could offer consumers a more reliable, understandable and concrete reasoning for saving energy.

4 VISUALISATION AND ANALYTICS

Energy management is one of the key domains where visual analytics can make an important contribution. With the introduction of smart meter technologies this statement applies as much to the high level energy demand and supply monitoring as to smarter controls and visual aids for householders. Technological advances in data visualisation offer real opportunities for research into energy consumption awareness with techniques that may provide personal views and interactive exploration of data. Rodgers and Bartram [7] encourage awareness and behavioural changes through tools and visualisations designed to make users aware of their energy use through non-intrusive and subtle visual stimuli.

Our research argues that the classification of energy-based variables could allow for variations in domestic energy consumption to be better understood. While data classification radically reduces data volumes and enables trends and clusters in large datasets to be identified with greater ease, they can also easily be misinterpreted. Some recent visual analytics research shows how exploratory data visualisation can be effective both during the clustering process and cluster decision stages (iVisCluster [2]) as well as for improving end-user understanding and overall comprehension (OAC Explorer [9]). This research highlights the benefits associated with classifying large datasets as well as taking steps to visualise the uncertainties that result from the clustering processes.

In our research we will use exploratory visualisation techniques to evaluate the use and benefits of our energy-based classifications. Requirements for the visualisations will be gathered at creativ-

ity workshops with E.ON staff during the ‘Visualising the Smart Home’ project.

5 RESEARCH STATUS

Having established a need for a geodemographic energy profile geovisualisation we are currently in the process of defining the classification system to be used, collecting datasets that may contribute and establishing visualisation requirements. Our poster will present some visual stimuli from the E.ON project as well as a description of our classifier and datasources to be used.

ACKNOWLEDGEMENTS

This PhD research is funded by a Vice Chancellor’s Scholarship from City University London and Betternest Ltd UK. We would also like to thank the contributors to the E.ON funded project.

REFERENCES

- [1] H. Allcott. Social Norms and Energy Conservation. *Journal of Public Economics*, 95(9–10):1082–1095, 2011.
- [2] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An Interactive VA System for Classification based on Supervised Dimension Reduction. In *VAST, 2010 IEEE Symposium on*, pages 27–34, Oct. 2010.
- [3] DECC. Average Household Electricity Consumption Data for the Standard Meter at LSOA level. (Available via decc.gov.uk), 2008.
- [4] A. Druckman and T. Jackson. Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model. *Energy Policy*, 36:3177–3192, 2008.
- [5] European Commission. Next Steps for Smart Grids: Europe’s Future Electricity System will Save Money and Energy. Technical report, Brussels, Apr. 2011.
- [6] Experian. Mosaic Public Sector. (Available via mimas.ac.uk), 2010.
- [7] J. Rodgers and L. Bartram. Exploring Ambient and Artistic Visualisation for Residential Energy Use Feedback. *IEEE TVCG*, 17(12):2489–2497, Dec. 2011.
- [8] A. D. Singleton and P. A. Longley. Geodemographics, Visualisation and Social Networks in Applied Geography. *Applied Geography*, 29(3):289–298, July 2009.
- [9] A. Slingsby, J. Dykes, and J. Wood. Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE TVCG*, 17(12):2545–2554, Dec. 2011.
- [10] D. Vickers and P. Rees. Creating the UK National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403, 2007.

A.3.1 Poster: IEEE VIS 2012, Seattle, USA

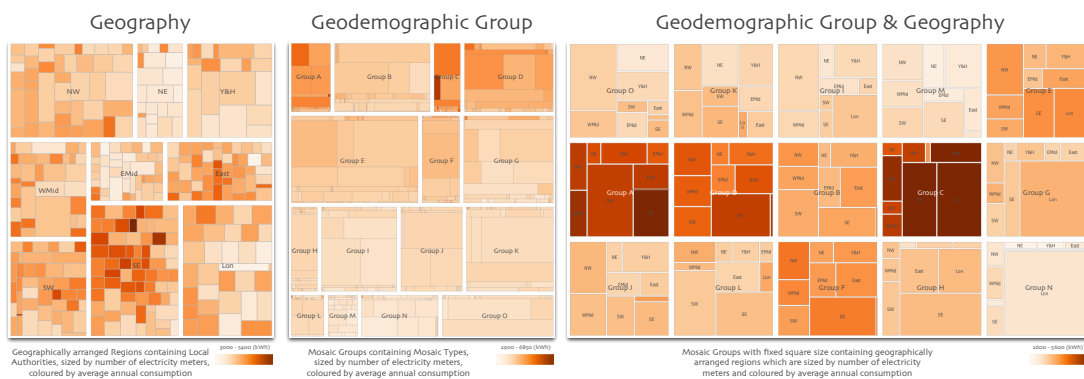
Visualising Variations in Household Energy Consumption

@sgeoviz
Sarah.Goodwin.1@city.ac.uk

Sarah Goodwin and Jason Dykes - giCentre, City University London, UK

Domestic energy consumption in the UK correlates with household disposable income, tenure, composition and urban/rural location[1], but the relationship between energy use and geodemographics has scarcely been investigated. We are analysing variations in energy consumption together with geography and geodemographics. A greater understanding of this complex relationship will benefit energy providers, local government and consumers as it will allow realistic comparisons and enable better targeting for services and schemes to encourage more sustainable energy use.

Visual Exploration: Household energy consumption varies by ...

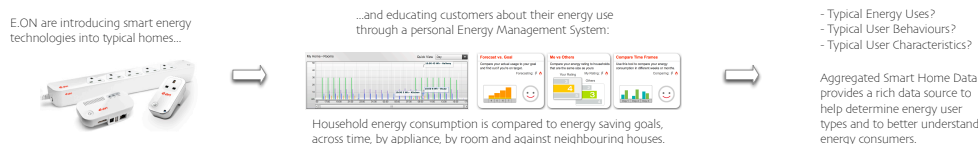


The exploratory analysis combines publicly available sub-national electricity consumption data for 2008 (based on ordinary electricity meters) from the Department of Climate Change (www.decc.gov.uk) combined with Experian's Mosaic Public Sector Classification 2010 available for academic research (www.mimas.ac.uk), which contains 15 demographic Groups and 69 Types. These visualisations were produced using HiDE software (www.gicentre.org/hide) [2].

The above visualisations show that both demographic group and geographic location correlate with energy consumption characteristics. Our research and industrial engagement identifies a need to better understand the patterns between population lifestyles and energy use, habits and behaviour.

Engagement with Industry: Visualising the Smart Home

The introduction of smart meter technology vastly increases the ability to better understand consumer energy use and behaviour. We are working in collaboration with the UK energy provider E.ON AG to explore new datasets which are becoming available following the initial adoption of smart home technologies. The 'Smart Home' provides many opportunities for visual analytics and data visualisation, with the ability to investigate energy consumption by aggregations of time, appliance and household characteristics. Our work with E.ON is helping to frame our research within the needs of the energy industry.



Our continued research aims to create a specifically defined neighbourhood energy-based classification that combines energy related datasets with relevant geodemographic variables. Engaging interactive geovisualisation techniques will be applied, developed and evaluated with industry experts to enable better interpretation and understanding of the final classification [3, 4].

[1] Druckman, A. & Jackson, T. (2008) 'Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model' Energy Policy, vol. 36, pp. 3177-3192.
[2] Slingsby, A., Dykes, J. & Wood, J. (2009) 'Configuring Hierarchical Layouts to Address Research Questions' IEEE Transactions on Visualization and Computer Graphics, 15(6), pp. 977-984.
[3] Choo, J., Lee, H., Kihm, J. and Park, H. (2010) 'VisClassifier: An Interactive Visual Analytics System for Classification based on Supervised Dimension Reduction' In VAST IEEE Symposium, pp. 27-34.
[4] Slingsby, A., Dykes, J. & Wood, J. (2011) 'Exploring Uncertainty in Geodemographics with Interactive Graphics' IEEE Transactions on Visualization and Computer Graphics, 17(12), pp. 2545-2554.



CITY UNIVERSITY
LONDON

gicentre.org
Conference Paper: openaccess.city.ac.uk/1294

A.4 Short Paper: GISRUUK 2012, Lancaster, UK

**Preprint paper for GIS Research UK 20th Annual Conference (GISRUUK 2012),
11 - 13 Apr 2012, Lancaster University, Lancaster UK.
Available at: <http://openaccess.city.ac.uk/895/>**

Geovisualization of Household Energy Consumption Characteristics

Sarah Goodwin¹ and Jason Dykes¹

¹giCentre, School of Informatics, City University London, EC1V OHB
Tel. 020 7040 8370
sarah.goodwin.1@city.ac.uk, j.dykes@city.ac.uk

Summary: A vast amount of quantitative data is available within the energy sector, however, there is limited understanding of the relationships between neighbourhoods, demographic characteristics and domestic energy consumption habits. We report upon research that will combine datasets relating to energy consumption, saving and loss with geodemographics to enable better understanding of energy user types. A novel interactive interface is planned to evaluate the performance of these energy-based classifications. The research aims to help local governments and the energy industry in targeting households and populations for new energy saving schemes and in improving efforts to promote sustainable energy consumption. Energy based neighbourhood classifications will also promote consumption awareness amongst domestic users. This poster describes the research methodology, data sources and visualization requirements.

KEYWORDS: Energy Consumption, Classification, Geodemographics, Visualization, Evaluation.

1. Introduction

Energy consumption is of growing interest to individuals, organizations and government due to EU energy consumption and carbon footprint reduction targets set for 2020 and 2050. In 2004 the household sector represented 27% of the UK's total carbon dioxide emissions and approximately 30% of total energy use (HM Government, 2006 in Druckman & Jackson, 2008). Achieving a large reduction at the domestic level is therefore imperative to meeting these targets. Over recent years there has been an increasing amount of research related to energy consumption, carbon reduction and potential energy saving opportunities; however, there is still limited knowledge of the relationship between energy consumption and measurable characteristics of population. Druckman and Jackson (2008) report that domestic fuel consumption in the UK is strongly related to disposable income levels with other highly influential factors being dwelling type, household composition, property tenure and rural/urban location. This work indicates that energy consumption patterns correlate to socio-economic and geographic characteristics and continued research in this field is needed in order to better target new low-carbon policies.

Within the energy industry there is a substantial amount of energy consumption data and the introduction of modern smart meters will increase this data quantity exponentially (Computer Weekly, 2009). Smart meter technology allows for consumption to be recorded at frequent intervals and communicates this information to both consumer and energy supplier allowing for near real-time feedback of energy use. Smart meters form a major component of 'Smart Grids', which are estimated to reduce annual EU household energy consumption by 10% and carbon dioxide emissions by 9% (European Commission, 2011). Many EU countries have started to introduce smart meter technology into households, with Italy reaching 85% household coverage in 2010 (Clastres, 2011). In 2009 the UK Government announced the intention to introduce smart meters into all households by 2020 (Faruqi *et al.*, 2010).

Smart meters are expected to greatly improve user awareness and allow for the regulation of energy consumption at the household level (Darby, 2010). A study of early adopters (Hargreaves *et al.*, 2010) reports improved awareness, but further studies are needed to identify whether changes in behaviour are long term and to understand the differences across household types. Traditionally household energy consumption feedback is provided through standard utility bills, which are usually vague,

uninformative and do not invite householders to think about their consumption patterns. It is rare for utility suppliers to provide benchmarks or comparison target groups for improving consumer awareness (Ehrhardt-Martinez *et al.*, 2010). Räsänen *et al.* (2008) acknowledge the need for neighbourhood level comparisons to provide consumers with understandable and concrete reasoning for saving energy as well as to encourage discussion of energy saving techniques amongst neighbours. Energy usage profiling and online visualization tools are now available for individuals to track their consumption over time, with some of these allowing for comparison at the neighbourhood or community level - such as iMeasure (Environmental Change Institute, 2011). While providing the opportunity to explore household energy consumption patterns in greater detail than the standard utility bill, these profiling tools are often time consuming and require some technical understanding in order to achieve a reduction in energy consumption. Ehrhardt-Martinez *et al.* (2010) review recent research and compare the success of different usage feedback schemes.

Electricity distribution deregulation has enabled electricity providers to formulate dedicated tariff types based on customer characteristics (Stephenson *et al.*, 2001). Energy providers and market analysts would benefit from consumption classifications as this would enable tariffs to be targeted based on typical consumer traits. Chicco *et al.* (2006, p.933) describe the need for electricity customer classifications for service providers:

“For the purpose of defining suitable tariff structures, the existing customer classifications based on the type of activity are scarcely correlated to the actual evolution of the electrical consumption and, as such, give poor information to the distribution providers”.

A new geographical clustering of energy characteristics at the neighbourhood level is necessary for energy companies, local governments and residents to allow realistic comparisons, understand complex consumption variations and enable better targeting of services and schemes to encourage more sustainable energy use.

2. Geodemographics and Energy Consumption

There is a body of relevant research correlating energy consumption with household or population variables (Semenik *et al.*, 1982); however, little research directly investigates the classification and evaluation of energy related variables with geodemographics. Druckman and Jackson (2008) compare energy consumption with the seven ONS Output Area Classification (OAC) Super Groups showing clear correlations with household disposable income and property tenancy. This draws parallels with other literature (Dillahunt & Mankoff, 2011; Dillahunt *et al.*, 2009) indicating that low-income families and tenant households have difficulties and additional barriers to reducing energy consumption.

Large multivariate dataset classification is ideal for residential energy consumption as previous research shows that human populations with similar characteristics and behaviours tend to cluster together. Some topical research by Chicco *et al.* (2003, 2006) evaluates techniques and methods for classifying characteristics of non-residential electricity use. In 2008, Experian introduced a data product ‘GreenAware’ (Experian, 2008) responding to demand to characterise populations based on energy behaviour. A case study of the use of this data by Haq and Owen (2009) demonstrates the potential for using population classifications to understand the geographical variations in energy consumption, however, the thematic map examples offered also highlight the difficulty in visualizing such abstract classified datasets. In classic thematic maps the areas of interest with the largest populations are frequently least visually salient due to the limited geographical area of the most densely populated areal units. Slingsby *et al.* (2011; 2010) show that well designed and novel visualization methods can be used to effectively visualize local and national multivariate datasets to overcome some of the problems associated with the kinds of thematic maps that are more routinely used.

3. Data Visualization and Energy Consumption

While data classification radically reduces data volumes and enables trends and clusters in large datasets to be identified with greater ease, they can be misinterpreted (Harris *et al.*, 2005). Data visualization can be both useful in helping gain access to and traction with such abstract but potentially informative information as well as providing insight into some of the detail lost during the classification process. OAC Explorer (Slingsby *et al.*, 2010; Slingsby *et al.*, 2011) shows that exploratory visualization methods can be effective in helping organizations understand local populations and their characteristics through geodemographic classifications. The public facing *placeSurvey* application (LSR, 2011) and related means of providing timely information for citizens demonstrate how visualization can be used to engage the public in exploratory analysis of information about local issues.

Technological advances in data visualization offer real opportunities for research into energy consumption awareness with techniques that may provide personal views and interactive exploration of energy data – potentially in real time. Recent research highlights new ideas to encourage awareness and behavioural changes through tools and visualizations designed to make the user aware of their current energy use through non-intrusive and subtle visual stimuli (Jönsson *et al.*, 2010; Rodgers *et al.*, 2011).

4. Research Plan and Status

The academic literature in the field highlights a continued need to classify UK energy user groups as well as provide the ability to explore such a classification through interactive visualization techniques. Our research therefore has two objectives:

- a. To create neighbourhood energy consumption classifications by combining datasets such as energy consumption, energy loss and saving potential with geodemographic variables
- b. To provide user groups such as energy suppliers, local government and citizens with the possibility to visualize this information through innovative and interactive geovisualization techniques that enable the data to be explored, understood, evaluated and acted upon.

The proposed classification and visualization of energy consumption related data will enable the private household energy market to be better understood, allow for energy profiles at the neighbourhood level and give local government and the energy industry better targets for potential energy saving schemes.

Having established a need for geodemographic energy profile geovisualization we are currently in the process of defining the classification system to be used, collecting datasets that may contribute to it and establishing visualization requirements. Our poster presents a description of our classifier and some initial visualization requirements that form the first stage in moving us towards our research objectives.

5. Acknowledgements

This work is funded by a Vice Chancellor's Scholarship from City University London and Betternest Ltd UK.

6. References

- Chicco, G, Napoli, R & Piglion, F 2003, "Application of Clustering Algorithms and Self Organising Maps to Classify Electricity Customers.," in *Power Tech Conference Proceedings*, Bologna, p. 7.

- Chicco, G, Napoli, R & Piglion, F 2006, "Comparisons Among Clustering Techniques for Electricity Customer Classification." *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 933-940.
- Clastres, C 2011, "Smart Grids: Another Step Towards Competition, Energy Security and Climate Change Objectives." *Energy Policy*, vol. 39, pp. 5399-5408.
- Computer Weekly 2009, "Smart Meters Multiply Data Loads." Retrieved November 22, 2011, from <http://www.computerweekly.com/news/2240089669/Smart-meters-multiply-data-loads>
- Darby, S 2010, "Smart Metering: What Potential for Householder Engagement?" *Building Research & Information*, vol. 38, no. 5, pp. 442-457.
- Dillahunt, T & Mankoff, J 2011, "In the Dark, Out in the Cold." *XRDS: Crossroads, The ACM Magazine for Students - Green Technologies*, vol. 17, no. 4, pp. 39-41.
- Dillahunt, T, Mankoff, J, Paulos, E & Fussell, S 2009, "It's Not All About 'Green': Energy Use in Low-Income Communities," in *Proceedings of the 11th international conference on Ubiquitous computing*, Ubicomp '09, ACM, New York, NY, USA, pp. 255-264.
- Druckman, A & Jackson, T 2008, "Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model." *Energy Policy*, vol. 36, pp. 3177-3192.
- Ehrhardt-Martinez, K, Donnelly, K & Laitner, J 2010, *Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities*, Retrieved November 22, 2011, from <http://www.aceee.org/research-report/e105>
- Environmental Change Institute 2011, "iMeasure: Home Energy and Carbon Monitoring Calculator." Retrieved November 30, 2011, from <http://www.imeasure.org.uk/>
- European Commission 2011, *Next Steps for Smart Grids: Europe's Future Electricity System will Save Money and Energy*, Brussels. Retrieved November 30, 2011, from http://ec.europa.eu/energy/gas_electricity/smartgrids/smartgrids_en.htm
- Experian 2008, "GreenAware: A Segmentation of Environmentally-Relevant Behaviours, Attitudes and Carbon Footprint." Retrieved November 19, 2011, from [http://www.experian.co.uk/assets/business-strategies/brochures/GreenAware_factsheet\[1\].pdf](http://www.experian.co.uk/assets/business-strategies/brochures/GreenAware_factsheet[1].pdf)
- Faruqui, A, Harris, D & Hledik, R 2010, "Unlocking the €53 Billion Savings from Smart Meters in the EU: How Increasing the Adoption of Dynamic Tariffs Could Make or Break the EU's Smart Grid Investment." *Energy Policy*, vol. 38, pp. 6222-6231.
- Haq, G & Owen, A 2009, "Green Streets The Neighbourhood Carbon Footprint of York." Retrieved January 25, 2011, from http://publicsector.experian.co.uk/Products/~/_media/CaseStudies/FinalGreenStreetsReportOct2009.aspx
- Hargreaves, T, Nye, M & Burgess, J 2010, "Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors." *Energy Policy*, vol. 38, no. 10, pp. 6111-6119.
- Harris, R, Sleight, P & Webber, R 2005, *Geodemographics: GIS and Neighbourhood Targeting*, Wiley-Blackwell.

- Jönsson, L, Broms, L & Katzeff, C 2010, "Watt-Lite: Energy Statistics made Tangible," in *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, ACM, New York, NY, USA, pp. 240–243.
- LSR Online (Leicestershire Statistics & Research) 2011, "PlaceSurvey." Retrieved November 30, 2011, from <http://www.lsr-online.org/placesurvey.html>
- Räsänen, T, Ruuskanen, J & Kolehmainen, M 2008, "Reducing Energy Consumption by Using Self-Organizing Maps to Create More Personalized Electricity Use Information." *Applied Energy*, vol. 85, no. 9, pp. 830-840.
- Rodgers, J, Bartram, L & Woodbury, R 2011, "Challenges in sustainable human-home interaction." *XRDS*, vol. 17, no. 4, pp. 42–46.
- Semenik, R, Belk, R & Painter, J 1982, "A Study of Factors Influencing Energy Conservation Behaviour." *Advances in Consumer Research*, vol. 09, pp. 306-312.
- Slingsby, A, Dykes, J & Wood, J 2011, "Exploring Uncertainty in Geodemographics with Interactive Graphics." *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2545 - 2554.
- Slingsby, A, Dykes, J, Wood, J & Radburn, R 2010, "OAC Explorer: Interactive Exploration and Comparison of Multivariate Socioeconomic Population Characteristics," in *Proceedings of the GIS Research UK*, pp. 167-174.
- Stephenson, P, Lungu, I, Paun, M, Silvas, I & Tupu, G 2001, "Tariff development for consumer groups in internal European electricity markets," in *Electricity Distribution, 2001. Part 1: Contributions. CIRED.*, IEE, Amsterdam.

7. Biography

Sarah Goodwin is a first-year PhD candidate at the giCentre, City University London. She has an academic and professional background in Geographical Information Science. After being awarded a distinction for her MSc at City University in 2007 and presenting her findings at GISRUk 2008 she was employed in Germany to analyse and map geographical variations in energy use characteristics.

Dr. Jason Dykes is Professor of Visualization at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.

A.4.1 Poster: GISRUK 2012, Lancaster, UK, - Won 'Best Poster'

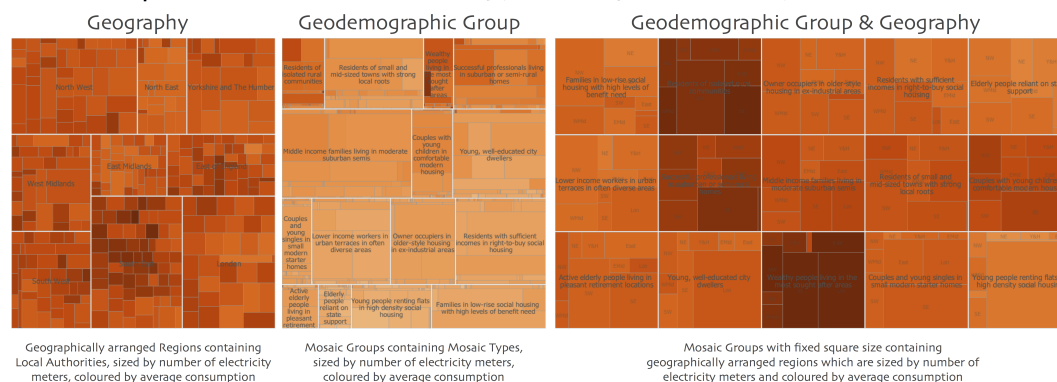
Geovisualization of Household Energy Consumption Characteristics

@sgeoviz
#sgeovizhide

Sarah Goodwin and Jason Dykes - giCentre, City University London

Domestic energy consumption correlates with household disposable income, tenure, composition and urban/rural location[1], but the relationship between energy use and geodemographics has scarcely been investigated. We are analysing variations in energy consumption together with geography and geodemographics. A greater understanding of this complex relationship at neighbourhood level will benefit local government, consumers and energy companies as it will allow realistic comparisons and enable better targeting for services and schemes to encourage more sustainable energy use.

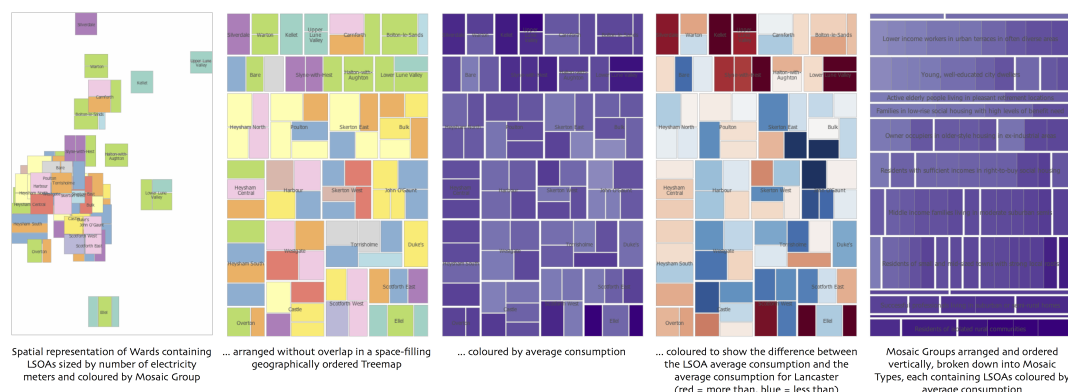
Visual Exploration: Household energy consumption varies by ...



This exploratory analysis combines publicly available 2008 electricity consumption data (using standard electricity meter data) from the Department of Climate Change (www.decc.gov.uk) with Experian's 2010 Mosaic Public Sector Classification available for academic research (www.mimas.ac.uk), which contains 15 demographic Groups and 69 Types. Our exploratory analysis was undertaken using HIDE software (www.giCentre.org/hide).

To improve accuracy when combining pre-aggregated datasets the highest resolution data available was used for the analysis (Lower Super Output Areas, ONS 2001 Census geography). Anomalies were found during the exploration, particularly in relation to unallocated electricity meters in the DeCC dataset which lead to discrepancies in household to meter count in certain areas. There is also a notable issue with the use of 2001 geographies as many LSOAs have become less homogenous since 2001, especially where large-scale regeneration has taken place. These issues will be addressed through our classification and visualization.

Lancaster Case Study: Transforming views to identify local patterns



Continued Research:

We aim to create a neighbourhood energy consumption classification that combines datasets related to energy consumption, saving and loss with geodemographics to enable better understanding of energy user types by applying and developing engaging interactive geovisualization techniques [2].

[1] Druckman, A & Jackson, T 2008, "Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model" Energy Policy, vol. 36, pp. 3177-3192.

[2] Slingsby, A., Dykes, J. & Wood, J. (2011). "Exploring Uncertainty in Geodemographics with Interactive Graphics". IEEE Transactions on Visualization and Computer Graphics, 17(12), pp. 2545-2554. doi: 10.1109/TVCG.2011.197 - <http://openaccess.city.ac.uk/437/>



CITY UNIVERSITY
LONDON

gicentre.org
Conference Paper: openaccess.city.ac.uk/895

A.5 Abstract: NACIS 2013, Greenville, USA

Abstract for NACIS (North American Cartographic Information Society's annual meeting) 2013,
Programme available: http://nacis.org/wp-content/uploads/2014/07/NACIS_2013_33rd.pdf

Visualizing Domestic Energy Consumption

Sarah Goodwin, Aidan Slingsby and Jason Dykes
giCentre, City University London

Abstract:

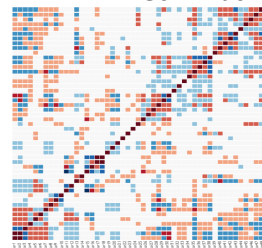
Growing populations and pressure to reduce worldwide CO₂ emissions has lead to an increased need to better understand the key drivers of domestic energy consumption. Despite energy consumption being a popular research topic in recent years, there is still a limited understanding on the relationship between energy use and measurable characteristics of the population. This presentation, of a UK-based PhD research project, reports on the exploration of this issue using data classification and geo-visualization techniques to identify geographic and demographic variations in domestic energy consumption characteristics. Such data classification enhances the ability to segment the domestic energy market, allowing for utility companies to group their consumers by typical traits and provide more tailored tariffs and services, while also enabling consumers to more reliably understand their household's usage against others.

14/10/2014

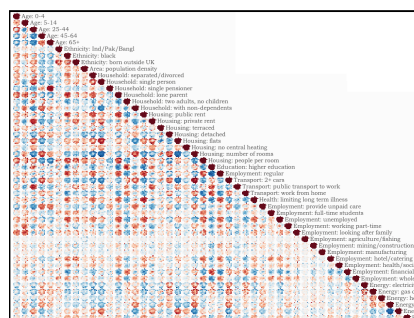
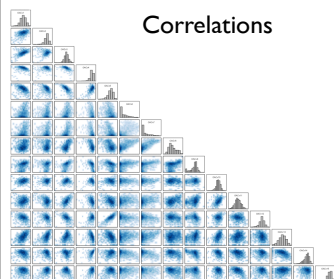
Variable Selection

- Gap in Research
- Correlations
- Distributions
- Spatial variability
- Many decisions – but how do these effect output?

Correlations



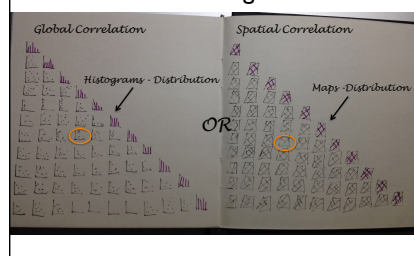
Correlations



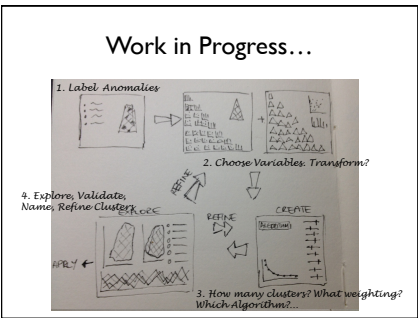
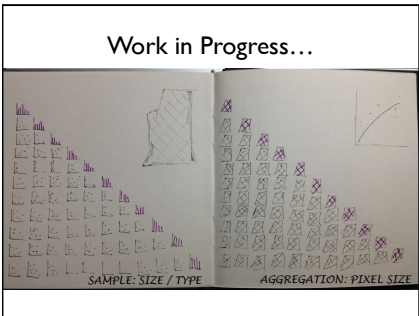
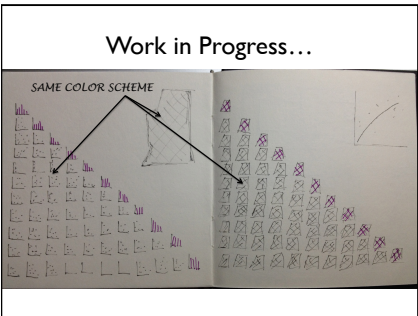
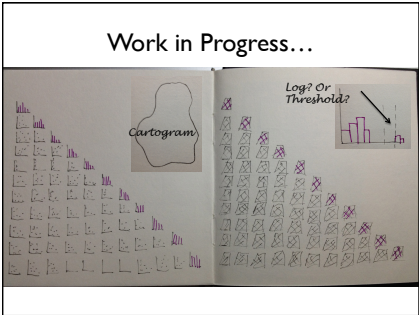
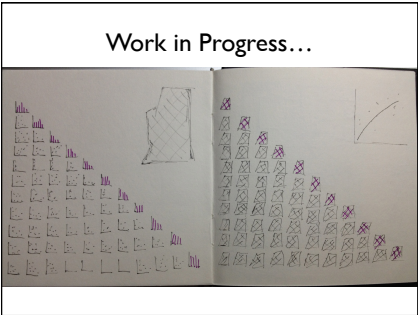
Goal

- To create a visual process to support all four stages and help:
- Choose & Transform Variables
 - Create Clusters
 - Explore Results
 - Refine Results
 - Name / Explain Clusters
 - Represent Uncertainty

Work in Progress...



14/10/2014



A.6 Paper: PhD Symposium 2012, Birmingham, UK

Paper for the 'Postgraduate symposium on household energy consumption, technology and efficiency' 6th June 2012, Energy, Society and Place Research Unit. University of Birmingham. Available online: <http://bit.ly/1ViFaKB>

Geovisualization of Household Energy Consumption Characteristics

Sarah Goodwin

giCentre, School of Informatics, City University London, EC1V 0HB

sarah.goodwin.1@city.ac.uk - 020 7040 8370

Summary: A vast amount of quantitative data is available within the energy sector, however, there is limited understanding of the relationships between neighbourhoods, demographic characteristics and domestic energy consumption habits. This research aims to combine datasets relating to energy consumption, saving and loss with geodemographics to enhance the understanding of energy user types. A novel interactive interface is planned to evaluate the performance of these energy-based classifications. The output produced aims to help local governments and the energy industry in targeting households and populations for new energy saving schemes and in improving efforts to promote sustainable energy consumption. Energy based neighbourhood classifications will also allow realistic comparisons amongst domestic users and help to promote consumption awareness.

1. Introduction

Energy consumption is of growing interest to individuals, organizations and government due to EU energy consumption and carbon footprint reduction targets set for 2020 and 2050. In 2004 the household sector represented 27% of the UK's total carbon dioxide emissions and approximately 30% of total energy use (HM Government, 2006 in Druckman & Jackson, 2008). Achieving a large reduction at the domestic level is therefore imperative to meeting these targets. Over recent years there has been an increasing amount of research related to energy consumption, carbon reduction and potential energy saving opportunities; however, there is still limited knowledge of the relationship between energy consumption and measurable characteristics of the population. Druckman and Jackson (2008) report that domestic fuel consumption in the UK is strongly related to disposable income levels with other highly influential factors being dwelling type, household composition, property tenure and rural/urban location. This research, along with others (such as Semenik et al., 1982; and more recently Dillahunty et al., 2009; Dillahunty & Mankoff, 2011), indicates that energy consumption patterns correlate to socio-economic and geographic characteristics and continued research in this field is needed in order to better understand the complex variations in consumption levels.

There is a growing need and demand for domestic energy users to be provided with a more transparent and detailed understanding of their energy usage. A neighbourhood level comparison based on users in the same category will offer consumers a more reliable, understandable and concrete reasoning for saving energy as well as potentially prompting discussion of energy saving techniques amongst neighbours (Räsänen et al., 2008). Energy providers and market analysts would also benefit from an energy-based user classification as this would enable tariffs to be better targeted based on typical consumer traits (Stephenson et al., 2001). A new geographical clustering of energy consumption data combined with household and population characteristics is therefore necessary to understand the complex consumption variations, allow realistic comparisons and facilitate better targeting of services and schemes to encourage more sustainable household energy use.

2. Geodemographics and Energy Consumption

Geodemographics is the 'analysis of people by where they live' (Sleight, 1997: 16 in Harris et al., 2005). Geodemographic classifications are developed using an area classification technique which combines geographic areas, such as postcodes, into groups based on the similarity of the associated characteristics such as age, employment status, average income, type of house and family size. Geodemographics is often used for business intelligence, customer profiling and targeted groups for direct marketing campaigns.

While there are a number of relevant research studies which correlate domestic energy consumption with household or population variables, such as household tenure or household income (Semenik et al., 1982), little research directly investigates the classification and evaluation of energy related variables with geodemographics. One such example, by Druckman and Jackson (2008), compares energy consumption data with the seven ONS Output Area Classification (OAC) Super Groups revealing clear correlations with household disposable income and property tenancy. Large multivariate dataset classification is ideal for residential energy consumption as studies show that populations with similar characteristics and behaviours tend to cluster together.

In 2008, Experian introduced the commercial data product 'GreenAware' (Experian, 2008) which allows businesses to target potential customers based on a carbon footprint measure and attitudes and behaviour towards green initiatives. A case study of the use of this classification by Haq and Owen (2009) demonstrates the potential for using population classifications to understand the geographical variations in energy consumption, however, the thematic map examples offered also highlight the difficulty in visualizing such abstract classified datasets. In classic thematic maps the areas of interest with the largest populations are frequently least visually salient due to the limited geographical area of the most densely populated areal units. Slingsby et al. (2011; 2010) show that well designed and novel visualization methods can be used to effectively visualize local and national multivariate datasets to overcome some of the problems associated with the kinds of thematic maps that are more routinely used.

3. Data Visualization and Energy Consumption

While data classification radically reduces data volumes and enables trends and clusters in large datasets to be identified with greater ease, classified data also has limitations and can be misinterpreted (Harris et al., 2005). The OAC Explorer application (Slingsby et al., 2010; 2011) shows that data visualization can effectively aid the representation, evaluation and user understanding of classified geodemographic data. Technological advances in data visualization offer real opportunities for research into energy consumption awareness with techniques that may provide personal views and interactive exploration of energy datasets. Recent research highlights new ideas to encourage awareness and behavioural changes through tools and visualizations designed to make the user aware of their current energy use through non-intrusive and subtle visual stimuli (Jönsson et al., 2010; Rodgers et al., 2011).

4. Research Plan and Status

The academic literature in the field highlights a continued need to classify UK energy user groups. This research study therefore has two objectives:

1. To create neighbourhood energy consumption classifications by combining datasets such as energy consumption, energy loss and saving potential with geodemographic variables
2. To provide user groups, such as energy suppliers, local government and citizens, with the possibility to visualize this information through innovative and interactive geovisualization techniques that enable the data to be explored, understood, evaluated and acted upon.

Some initial data analysis and visualizations created with openly available energy consumption and geodemographic datasets reveal both demographic group patterns and geographical variations. Research is currently underway to source, collect, combine and correlate relevant datasets, define the classification system to be used as well as establish the necessary visualization requirements.

References

- Dillahunt, T. et al., 2009. It's Not All About "Green": Energy Use in Low-Income Communities. In *Proceedings of the 11th international conference on Ubiquitous computing*. Ubicomp '09. New York, NY, USA: ACM, pp. 255–264.

- Dillahunt, T. & Mankoff, J., 2011. In the Dark, Out in the Cold. *XRDS: Crossroads, The ACM Magazine for Students - Green Technologies*, 17(4), pp.39–41.
- Druckman, A. & Jackson, T., 2008. Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model. *Energy Policy*, 36, pp.3177–3192.
- Experian, 2008. GreenAware: A Segmentation of Environmentally-Relevant Behaviours, Attitudes and Carbon Footprint. Available at: [http://www.experian.co.uk/assets/business-strategies/brochures/GreenAware_factsheet\[1\].pdf](http://www.experian.co.uk/assets/business-strategies/brochures/GreenAware_factsheet[1].pdf) [Accessed November 19, 2011].
- Haq, G. & Owen, A., 2009. Green Streets The Neighbourhood Carbon Footprint of York. Available at: <http://publicsector.experian.co.uk/Products/~media/CaseStudies/FinalGreenStreetsReportOct2009.ashx> [Accessed January 25, 2011].
- Harris, R., Sleight, P. & Webber, R., 2005. *Geodemographics: GIS and Neighbourhood Targeting*, Wiley-Blackwell.
- Jönsson, L., Broms, L. & Katzeff, C., 2010. Watt-Lite: Energy Statistics made Tangible. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. DIS '10. New York, NY, USA: ACM, pp. 240–243.
- Räsänen, T., Ruuskanen, J. & Kolehmainen, M., 2008. Reducing Energy Consumption by Using Self-Organizing Maps to Create More Personalized Electricity Use Information. *Applied Energy*, 85(9), pp.830–840.
- Rodgers, J., Bartram, L. & Woodbury, R., 2011. Challenges in sustainable human-home interaction. *XRDS*, 17(4), pp.42–46.
- Semenik, R., Belk, R. & Painter, J., 1982. A Study of Factors Influencing Energy Conservation Behaviour. *Advances in Consumer Research*, 09, pp.306–312.
- Slingsby, A. et al., 2010. OAC Explorer: Interactive Exploration and Comparison of Multivariate Socioeconomic Population Characteristics. In *Proceedings of the GIS Research UK*. 18th Annual Conference GISRUUK 2010. pp. 167–174.
- Slingsby, A., Dykes, J. & Wood, J., 2011. Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp.2545 – 2554.
- Stephenson, P. et al., 2001. Tariff development for consumer groups in internal European electricity markets. In *Electricity Distribution, 2001. Part 1: Contributions*. CIREN. 16th International Conference and Exhibition on (IEE Conf. Publ No. 482). Amsterdam: IEE.

Biography

Sarah Goodwin is a first-year PhD candidate at the giCentre, City University, under the supervisor of Professor Jason Dykes. She has an academic and professional background in Geographical Information Science. After being awarded a distinction for her MSc at City University in 2007 she was employed in Germany to analyse and map geographical variations in energy use characteristics.

B

Additional Material

The following pages contain additional material relevant to the research including consent forms, questionnaires, photos of the workshop, requirements and design generation, details of the variable sources and brainstorming visual designs for the framework.

B.1 Smart Home Project Consent Form

Participant Consent Form

Project Title: *Visualising the Smart Home: Creative Engagement with Customer Data*

I agree to take part in the City University London research project named above and undertaken in conjunction with E.ON Energy. I have had the project explained to me.

Data Protection

I agree to City University London recording and processing this information about me. I understand that this information will be used only for the purposes set out in this statement and my consent is conditional on the University complying with its duties and obligations under the Data Protection Act 1998.

Withdrawal from Study

I understand that my participation is voluntary and that I can withdraw at any stage without being penalised or disadvantaged in any way.

Photography and Audio

I understand that during this workshop photographs and audio recordings will be made. I understand that any such photography or audio will be used for data analysis and that selected images may be used for disseminating any findings in academic publications or project reports. I understand that if I choose not to be photographed or recorded I will not be penalised or disadvantaged in any way but that I will be asked to wear a marker identifying this choice.

☐ I give my consent for City University to use any photography taken during the workshop in which I may be present

☐ I give my consent for City University to use any audio recordings taken during the workshop in which I may be present

Name:(please print)

Signature: Date:.....

B.2 Photos from the Requirements Workshop



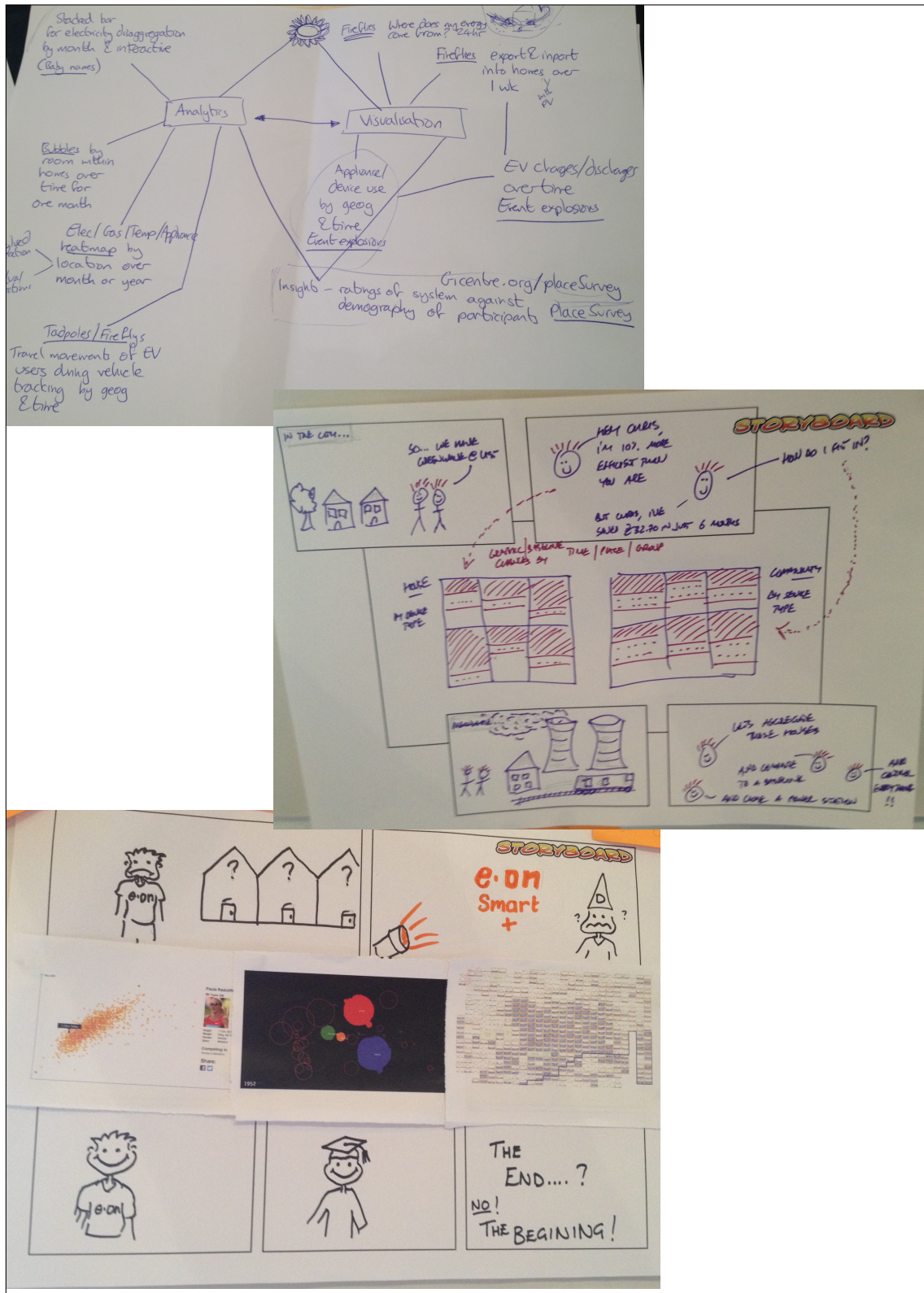
B.3 Aspirations and Storyboards from Workshop

A review of the statements and informal requirements gathered during the creativity workshop with E.ON

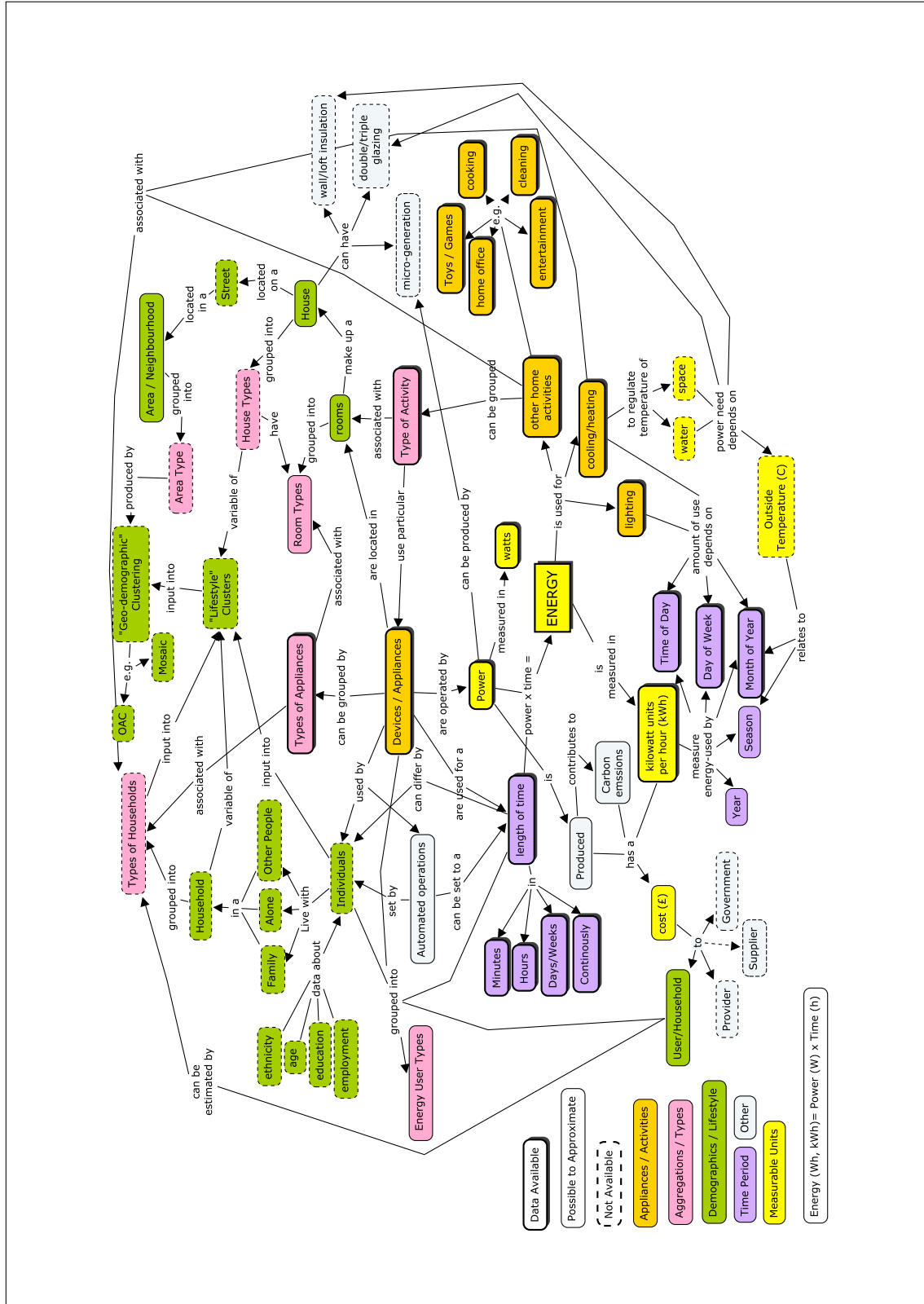
Key to Symbols used:
 ✓✓ Definitely Feasible / Met
 ✓ Feasible / Met to a lesser degree
 ! Partly Feasible / Partly Met
 * Could be met after feedback
 ✗ Not feasible / Not met

ACTIVITY	STATEMENT	Met	FEASIBLE?	Demand Horizons	Ownership Groups	Consumption Signatures	Smart Home Heatlines	Comment
<i>Would like to</i> See	Everything in 3 clicks: Simplicity / easy to interact with	-	✓✓	✓✓	✓✓	✓✓	✓✓	
See	Accuracy, but precise data (accuracy is important)	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
See	Who factor - need to see things in a different way which engages stakeholders	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
See	Who is using what and when - understand more about their usage.	Y	✓✓	✓✓	✓✓	✓✓	✓✓	who is difficult
See	Demand side management and time of use tariffs.	Y	✓	✓	✓	✓	✓	
See	Exactly where the energy (elec/gas/heat) is used / generated in the home	Y	✓	✓	✓	✓	✓	
See	Show the "effect" of technology (cost / carbon / comfort)	-	✓	✓	✓	✓	✓	exactly where needs more data
See	What the householder sees: so you know how and why they are operating	-	✓	✓	✓	✓	✓	why and who is difficult
See	How comfortable people are (secure, peace of mind).	-	✓	✓	✓	✓	✓	
<i>Would like to</i> Know	What device - when is it used, how long is it used for?	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	What does an "average" home do with energy [baseline]	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	How to visualize disaggregated energy use.	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	Who is using what and when - understand more about their usage.	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	Why do people use energy?	-	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	Who, what, where, when, why?	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	Where is electricity/gas generated or used in homes.	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	How electricity and gas use changes with age (lifestyle, life cycle, group, segment).	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	More about the value to E.ON of data (of data/insight and visualisations)	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	a baseline to use	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	The "inside out of energy consumption" in a real home	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	"Standby vs On - what's really going on with energy consumption?"	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	How to optimise Home Consumption / generation	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	How to show to the Business "how much, where, when and who"	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	How lifestyle links to energy demand	Y	✓✓	✓✓	✓✓	✓✓	✓✓	
Know	Who is involved, what is going on, how are they managing energy, why are they managing it in this way?	-	✓	✓	✓	✓	✓	
Know	How comfortable people are (secure, peace of mind).	-	✓	✓	✓	✓	✓	Customer Feedback
Know	Customer comparisons (baseline needed) to let people know what effect smart technology has had on them.	!	✓	✓	✓	✓	✓	Customer Feedback
Know	How to optimise the customer experience	-	✓	✓	✓	✓	✓	More data needed
Know	Customer drivers for connected energy	-	✓	✓	✓	✓	✓	Task 4 / Next of Project
Know	Segmented energy use by individual within household.	-	✓	✓	✓	✓	✓	More data needed
<i>Would like to</i> Do	Easily interact (3 clicks) & simply engage with the data	Y	✓✓	✓✓	✓✓	✓✓	✓✓	Customer Feedback
Do	Inform trading (through understanding customers)	Y	✓✓	✓✓	✓✓	✓✓	✓✓	Customer Feedback
Do	Manage Smart Homes effectively - so we can continue building on it	Y	✓✓	✓✓	✓✓	✓✓	✓✓	More data / Analysis needed
Do	Support the grid [using data to manage / even out demand]	Y	✓✓	✓✓	✓✓	✓✓	✓✓	More analysis / data needed
Do	Give useful tips for savings to users - based on data	Y	✓✓	✓✓	✓✓	✓✓	✓✓	In future
Do	understand how the data relates to the people in the houses	Y	✓	✓	✓	✓	✓	More analysis / data needed
Do	Provide it at right cost - the savings to customer outweighs the cost (need the data to prove this).	-	✓	✓	✓	✓	✓	Customer Feedback / More data
Do	Engage customers with/about energy	Y	✓	✓	✓	✓	✓	
Do	Analyse data for insights that are to do with more than energy used (but peace of mind, comfort, security etc.)	-	✓	✓	✓	✓	✓	
Do	Connect everything (appliances, houses, technology) together.	-	✓	✓	✓	✓	✓	
Do	Control everything! - individual devices, groups of devices, flexible demand.	-	✓	✓	✓	✓	✓	
Do	"Set & Forget" technology for customers	-	✓	✓	✓	✓	✓	
Do	Control energy in homes without conflicting with customer preferences.	-	✓	✓	✓	✓	✓	
Do	Engage with customers in a long term way "through" energy.	-	✓	✓	✓	✓	✓	

B.3. ASPIRATIONS AND STORYBOARDS FROM WORKSHOP



B.4 Concept Map of Energy Analysis Possibilities

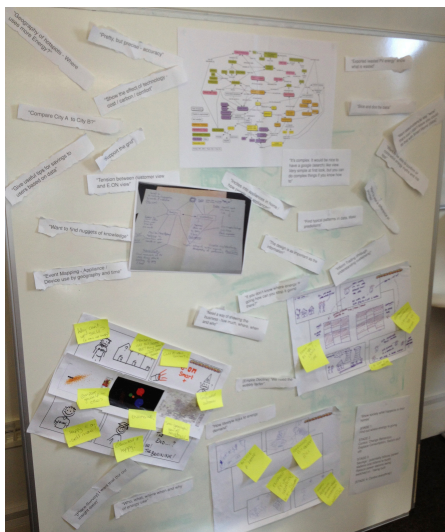


B.5 Photos from the Internal Design Workshop

Internal giCentre Smart Home Design Workshop– 22/11/2012

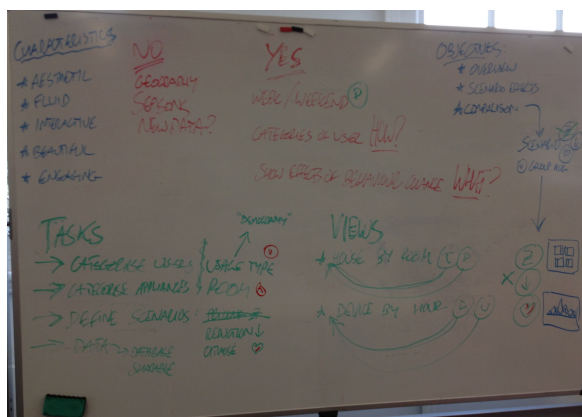
Preparation:

- Presentation of the **project overview** and background context of smart energy grid
- And a breakdown of the two **data sets**: Trial data and Model.
- **Requirements** from the workshop with E.ON – **Quotes, Storyboards and Key themes and ideas** from the different activities (all placed on a white board).
- Content diagram linking ideas together



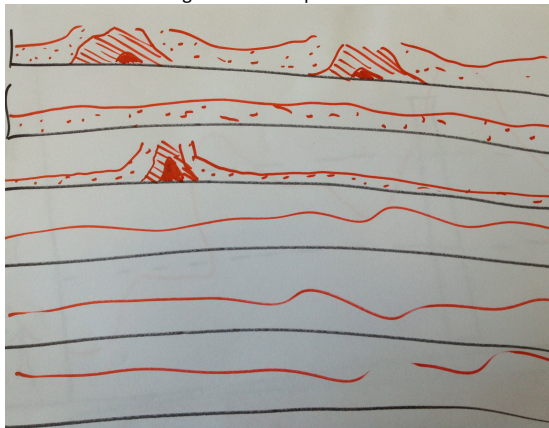
Brainstorming

We allowed time to discuss the data, read the quotes and ideas and investigate the options available. We discussed possibilities for visualisation in relation to time and data available as well as our expertise. We identified some key items to move forward with in our designs:





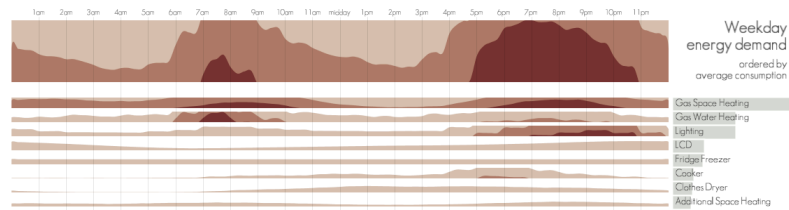
Working in pairs, to brainstorm together, we came up with four possible ideas for visualisations which related to the requirements and the project goals. Sketches were created for some.. e.g. Horizon Graph sketch



Initial Ideas

1. **Horizon Graphs** – Show total consumption and total by appliances. It will show patterns by appliance over time and their contribution to the peaks.
2. **Grouping /filtering data interactively** – analytics of appliances by type
3. **Optimization Scenarios** – compare options against original
4. **Interactive Data Exploration** of E.ON Trial Data.

B.6 Questionnaire of Appropriateness



HorizEon

Please rate each of the following statements from 1 to 6 based on the amount you agree with it, with 1 being strongly agree and 6 being strongly disagree.

"This visualization technique would enable me to understand how different appliances contribute to the peaks in energy demand throughout the day."

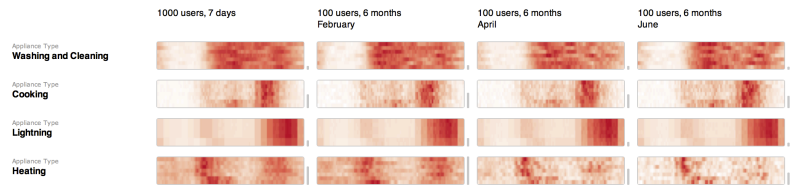
Strongly Agree 1 2 3 4 5 6 Strongly Disagree

"This visualization technique would be an engaging way to share data about how much energy is consumed, where it is consumed and when it is consumed."

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

"This visualization technique would enable me to support a dynamic electricity grid by simulating shifts in energy demand or time of consumption."

Strongly Agree 1 2 3 4 5 6 Strongly Disagree



SignatureApp

Please rate each of the following statements from 1 to 6 based on the amount you agree with it, with 1 being strongly agree and 6 being strongly disagree.

“This visualization technique would be an effective method with which to view large amounts of energy consumption data.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

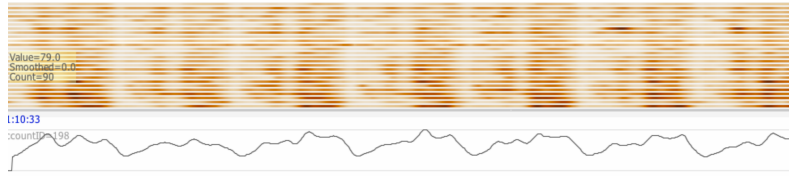
“This visualization technique would help me to understand the effects of seasonality on different appliances or groups of appliances.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

“This visualization technique would help me to find out the different times at which appliances or groups of appliances are used.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

B.6. QUESTIONNAIRE OF APPROPRIATENESS



SmartExplorer

Please rate each of the following statements from 1 to 6 based on the amount you agree with it, with 1 being strongly agree and 6 being strongly disagree.

"This visualization technique would help me to identify possible errors in the data."

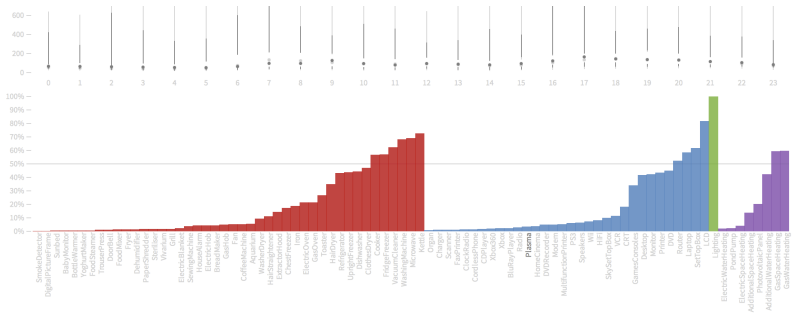
Strongly Agree 1 2 3 4 5 6 Strongly Disagree

"This visualization technique would be an effective way to show the large amounts of raw energy consumption data generated by individual households."

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

"This visualization technique would enable me to compare the energy consumption of individual households or groups of households."

Strongly Agree 1 2 3 4 5 6 Strongly Disagree



SmartGroups

Please rate each of the following statements from 1 to 6 based on the amount you agree with it, with 1 being strongly agree and 6 being strongly disagree.

“These visualization techniques would help me to gain insights into which appliances customers own and how they use them.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

“These visualization techniques would enable me to relate different components of the data to each other.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

“These visualization techniques would enable me to group the energy consumption of different customers based upon the appliances they have.”

Strongly Agree 1 2 3 4 5 6 Strongly Disagree

B.7 Generating Geodemographics Requirements



APPENDIX B. ADDITIONAL MATERIAL

B.8 167 variables considered for OAC 2011 from Gale, 2014, pp.224

OAC 2011 - Variable Code, Variable Name and Variable Domains:

Demographic (Red), Household Composition (Blue), Housing (Green), Socio-Economic (Purple) and Employment (Orange).

Code	Variable Name		
u001	Males	u051	Households that only contain Persons aged over 16 who are living in a couple: Married
u002	Females	u052	Households that only contain Persons aged over 16 who are living in a couple: Cohabiting (opposite-sex)
u003	Persons living in a household	u053	Households that only contain Persons aged over 16 who are living in a couple: In a registered same-sex civil partnership or cohabiting (same-sex)
u004	Persons living in a communal establishment	u054	Households that only contain Persons aged over 16 who are not living in a couple: Single (never married or never registered a same-sex civil partnership)
u005	Area size (in hectares)	u055	Households that only contain Persons aged over 16 who are not living in a couple: Married or in a registered same-sex civil partnership
u006	Number of persons per hectare	u056	Households that only contain Persons aged over 16 who are not living in a couple: Separated (but still legally married or still legally in a same-sex civil partnership)
u007	Persons aged 0 to 4	u057	Households that only contain Persons aged over 16 who are not living in a couple: Divorced or formerly in a same-sex civil partnership which is now legally dissolved
u008	Persons aged 5 to 9	u058	Households that only contain Persons aged over 16 who are not living in a couple: Widowed or surviving partner from a same-sex civil partnership
u009	Persons aged 10 to 14	u059	One person households: Aged 65 and over
u010	Persons aged 15 to 19	u060	One person households: Other
u011	Persons aged 20 to 24	u061	One family households: All aged 65 and over
u012	Persons aged 25 to 29	u062	One family households: Married or same-sex civil partnership couple with no children
u013	Persons aged 30 to 44	u063	One family households: Married or same-sex civil partnership couple with dependant children
u014	Persons aged 45 to 59	u064	One family households: Married or same-sex civil partnership couple with non-dependant children
u015	Persons aged 60 to 64	u065	One family households: Cohabiting couple with no children
u016	Persons aged 65 to 74	u066	One family households: Cohabiting couple with dependant children
u017	Persons aged 75 to 84	u067	One family households: Cohabiting couple with non-dependant children
u018	Persons aged 85 to 89	u068	One family households: Lone parent with dependant children
u019	Persons aged 90 and over	u069	One family households: Lone parent with non-dependant children
u020	Mean age	u070	Other household types: With dependant children
u021	Median age	u071	Other household types: All full-time students
u022	Persons aged over 16 who are single	u072	Other household types: All aged 65 and over
u023	Persons aged over 16 who are married	u073	Other household types: Other
u024	Persons aged over 16 who are in a registered same-sex civil partnership	u074	Households with no adults in employment: With dependant children
u025	Persons aged over 16 who are separated	u075	Households with no adults in employment: No dependant children
u026	Persons aged over 16 who are divorced or formerly in a same-sex civil partnership which is now legally dissolved	u076	Households with lone parent in part-time employment
u027	Persons aged over 16 who are widowed or a surviving partner from a same-sex civil partnership	u077	Households with lone parent in full-time employment
u028	Persons who are white British and Irish	u078	Households with lone parent not in employment
u029	Persons who are other white	u079	One person ethnic household
u030	Persons who have mixed ethnicity or are from multiple ethnic groups	u080	Household members all have the same ethnic group
u031	Persons who are Asian/Asian British: Indian	u081	Households with different ethnic groups between the generations only
u032	Persons who are Asian/Asian British: Pakistani	u082	Households with different ethnic groups within partnerships (whether or not different ethnic groups between generations)
u033	Persons who are Asian/Asian British: Bangladeshi	u083	Households with any other combination of multiple ethnic groups
u034	Persons who are Asian/Asian British: Chinese	u084	Household spaces with at least one usual resident
u035	Persons who are Asian/Asian British: Other	u085	Household spaces with no usual residents
u036	Persons who are Black/African/Caribbean/Black British	u086	Households who live in a detached house or bungalow
u037	Persons who are Arab or are from another ethnic group	u087	Households who live in a semi-detached house or bungalow
u038	Persons who are Christian	u088	Households who live in a terrace or end-terrace house
u039	Persons who are from another religion	u089	Households who live in a flat
u040	Persons who have no religion	u090	Households who live in a caravan or other mobile or temporary structure
u041	Persons who did not state their religion	u091	Households who own or have shared ownership of property
u042	Persons whose country of birth is the United Kingdom	u092	Households who are private renting
u043	Persons whose country of birth is Ireland	u093	Households who are social renting
u044	Persons whose country of birth is in the old EU (pre 2004 accession countries)	u094	Households who are living rent free
u045	Persons whose country of birth is in the new EU (post 2004 accession countries)	u095	Households who have two or more rooms than required
u046	Persons whose country of birth is not the UK, Ireland or EU countries	u096	Households who have one more room than required
u047	Persons whose main language is English or their main language is not English but can speak English very well	u097	Households who have the required number of rooms
u048	Persons whose main language is not English but can speak English well	u098	Households who have one fewer room than required
u049	Persons whose main language is not English and cannot speak English well	u099	Households who have two fewer or less rooms than required
u050	Persons whose main language is not English and cannot speak English	u100	Households with up to 0.5 persons per room
		u101	Households with over 0.5 and up to 1.0 persons per room
		u102	Households with over 1.0 and up to 1.5 persons per room
		u103	Households with over 1.5 persons per room

B.8. 167 VARIABLES CONSIDERED FOR OAC 2011 FROM GALE, 2014,PP.224

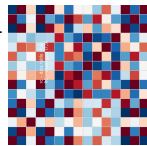

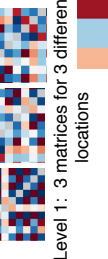
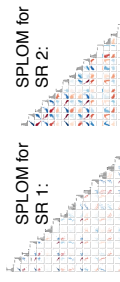
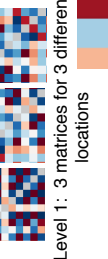

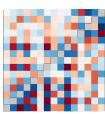
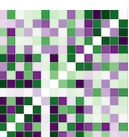
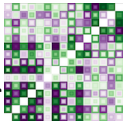

OAC 2011 - Variable Code, Variable Name and Variable Domains:
Demographic (Red), Household Composition (Blue), Housing (Green), Socio-Economic (Purple) and Employment (Orange).

u104	Day-to-day activities limited a lot or a little Standardised Illness Ratio	u148	Employed persons aged between 16 and 74 industry: Transport and storage
u105	Persons in very good health	u149	Employed persons aged between 16 and 74 industry: Accommodation and food service activities
u106	Persons in good health	u150	Employed persons aged between 16 and 74 industry: Information and communication
u107	Persons in fair health	u151	Employed persons aged between 16 and 74 industry: Financial and insurance activities
u108	Persons in bad health	u152	Employed persons aged between 16 and 74 industry: Real estate activities
u109	Persons in very bad health	u153	Employed persons aged between 16 and 74 industry: Professional, scientific and technical activities
u110	Persons providing unpaid care	u154	Employed persons aged between 16 and 74 industry: Administrative and support service activities
u111	Persons aged over 16 who have no qualifications	u155	Employed persons aged between 16 and 74 industry: Public administration and defence; compulsory social security
u112	Persons aged over 16 whose highest level of qualification is Level 1, Level 2 or Apprenticeship	u156	Employed persons aged between 16 and 74 industry: Education
u113	Persons aged over 16 whose highest level of qualification is Level 3 qualifications	u157	Employed persons aged between 16 and 74 industry: Human health and social work activities
u114	Persons aged over 16 whose highest level of qualification is Level 4 qualifications and above	u158	Employed persons aged between 16 and 74 industry: Other industry
u115	Persons aged over 16 who are schoolchildren or full-time students	u159	Employed persons aged between 16 and 74 occupation: Managers, directors and senior officials
u116	Households with no cars or vans	u160	Employed persons aged between 16 and 74 occupation: Professional occupations
u117	Households with 1 car or van	u161	Employed persons aged between 16 and 74 occupation: Associate professional and technical occupations
u118	Households with 2 or more cars or vans	u162	Employed persons aged between 16 and 74 occupation: Administrative and secretarial occupations
u119	Persons aged between 16 and 74 who work mainly at or from home	u163	Employed persons aged between 16 and 74 occupation: Skilled trades occupations
u120	Persons aged between 16 and 74 who use public transport to get to work	u164	Employed persons aged between 16 and 74 occupation: Caring, leisure and other service occupations
u121	Persons aged between 16 and 74 who use private transport to get to work	u165	Employed persons aged between 16 and 74 occupation: Sales and customer service occupations
u122	Persons aged between 16 and 74 who walk, cycle or use an alternative method to get to work	u166	Employed persons aged between 16 and 74 occupation: Process, plant and machine operatives
u123	Persons aged between 16 and 74 who are economically active: Part-time employees	u167	Employed persons aged between 16 and 74 occupation: Elementary occupations
u124	Persons aged between 16 and 74 who are economically active: Full-time employees		
u125	Persons aged between 16 and 74 who are economically active: Self-employed		
u126	Persons aged between 16 and 74 who are economically active: Unemployed		
u127	Persons aged between 16 and 74 who are economically active: Full-time student		
u128	Persons aged between 16 and 74 who are economically inactive: Retired		
u129	Persons aged between 16 and 74 who are economically inactive: Student (including full-time students)		
u130	Persons aged between 16 and 74 who are economically inactive: Looking after home or family		
u131	Persons aged between 16 and 74 who are economically inactive: Long-term sick or disabled		
u132	Persons aged between 16 and 74 who are economically inactive: Other		
u133	Persons aged between 16 and 24 who are unemployed		
u134	Persons aged between 50 and 74 who are unemployed		
u135	Persons aged between 16 and 74 who have never worked		
u136	Persons aged between 16 and 74 who are long-term unemployed		
u137	Employed persons aged between 16 and 74: Part-time working 15 hours or less		
u138	Employed persons aged between 16 and 74: Part-time working 16 to 30 hours		
u139	Employed persons aged between 16 and 74: Full-time working 31 to 48 hours		
u140	Employed persons aged between 16 and 74: Full-time working 49 or more hours		
u141	Employed persons aged between 16 and 74 industry: Agriculture, forestry and fishing		
u142	Employed persons aged between 16 and 74 industry: Mining and quarrying		
u143	Employed persons aged between 16 and 74 industry: Manufacturing		
u144	Employed persons aged between 16 and 74 industry: Electricity, gas, steam and air conditioning supply		
u145	Employed persons aged between 16 and 74 industry: Water supply; sewerage, waste management and remediation activities		
u146	Employed persons aged between 16 and 74 industry: Construction		
u147	Employed persons aged between 16 and 74 industry: Wholesale and retail trade; repair of motor vehicles and motor cycles		

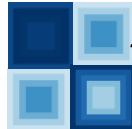


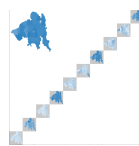


B.9 Final 78 Variables used in Framework Prototype

Variable	Name	Domain Name	OAC2001	OAC2011	Data Type	Denominator	CensusTable
1	Central Heating: Electricity	Energy			Percentage	Total Households	KS403EW
2	Energy: Annual Electricity Consumption	Energy			Average	Total Households	DECC
3	Energy: in Fuel Poverty	Energy			Percentage	Total Households	DECC
4	Central Heating: Gas	Energy			Percentage	Total Households	KS403EW
5	Energy: Annual Gas Consumption	Energy			Average	Total Households	DECC
6	Central Heating: None	Energy	v21		Percentage	Total Households	KS403EW
7	Central Heating: Other (Wood, Coal, Oil)	Energy			Percentage	Total Households	KS403EW
8	Household: No English Language	Demographic		v23	Percentage	Total Households	KS206EW
9	Population: Aged 0 - 4	Demographic	v1	v1	Percentage	Total Population	KS102EW
10	Population: Aged 5 - 14	Demographic	v2	v2	Percentage	Total Population	KS102EW
11	Population: Aged 25 - 44	Demographic	v3	v3	Percentage	Total Population	KS102EW
12	Population: Aged 45 - 64	Demographic	v4	v4	Percentage	Total Population	KS102EW
13	Population: Aged 65+	Demographic	v5		Percentage	Total Population	KS102EW
14	Population: Aged 65 to 89	Demographic		v5	Percentage	Total Population	KS102EW
15	Population: Aged 90 and over	Demographic		v6	Percentage	Total Population	KS102EW
16	Born: in new (post 2004) EU Countries	Demographic		v21	Percentage	Total Population	KS204EW
17	Born: in old (pre 2004) EU Countries	Demographic		v22	Percentage	Total Population	KS204EW
18	Born: Outside the UK	Demographic	v8		Percentage	Total Population	KS201EW
19	Born: United Kingdom and Ireland	Demographic		v20	Percentage	Total Population	KS204EW
20	Ethnicity: Arab or other ethnic groups	Demographic		v19	Percentage	Total Population	KS201EW
21	Ethnicity: Asian: Bangladeshi	Demographic		v16	Percentage	Total Population	KS201EW
22	Ethnicity: Black African, Caribbean or Other	Demographic	v7	v18	Percentage	Total Population	KS201EW
23	Ethnicity: Asian: Chinese and Other	Demographic		v17	Percentage	Total Population	KS201EW
24	Ethnicity: Asian: Indian	Demographic		v14	Percentage	Total Population	KS201EW
25	Ethnicity: Indian, Pakistani or Bangladeshi	Demographic	v6		Percentage	Total Population	KS201EW
26	Ethnicity: Mixed ethnic group	Demographic		v13	Percentage	Total Population	KS201EW
27	Ethnicity: Asian: Pakistani	Demographic		v15	Percentage	Total Population	KS201EW
28	Ethnicity: White	Demographic		v12	Percentage	Total Population	KS201EW
29	Urban: Population Density	Demographic	v9	v7	Ratio		KS101EW
30	Household: with no children	Household Composition	v14	v25	Percentage	Total Households	KS105EW
31	Household: with non-dependant children	Household Composition	v15	v24	Percentage	Total Households	KS105EW
32	House: Communal Establishment	Household Composition		v8	Percentage	Total Households	KS101EW
33	Household: Full-time student(s)	Household Composition		v26	Percentage	Total Households	KS105EW
34	Household: Lone Parent	Household Composition	v13		Percentage	Total Households	KS105EW
35	Status: Married or civil partnership	Household Composition		v10	Percentage	Total Population (16 and over)	KS105EW
36	Status: Separated/Divorced	Household Composition	v10	v11	Percentage	Total Population (16 and over)	KS105EW
37	Status: Single	Household Composition		v9	Percentage	Total Population	KS105EW
38	Household: Single person (not pensioner)	Household Composition	v11		Percentage	Total Households	KS105EW
39	Household: Single pensioner	Household Composition	v12		Percentage	Total Households	KS105EW
40	House: Detached	Housing Type	v19	v27	Percentage	Total Households	KS401EW
41	House: Flat/Apartment	Housing Type	v20	v30	Percentage	Total Households	KS401EW
42	Overcrowding: Average house size	Housing Type	v22		Average	Total Households	KS403EW
43	Tenure: Owned or Shared Ownership	Housing Type		v31	Percentage	Total Households	KS402EW
44	Overcrowding: People per room	Housing Type	v23		Average		KS403EW
45	Average number of rooms	Housing Type			Average	Total Households	KS403EW
46	Tenure: Rent (Private)	Housing Type	v17	v33	Percentage	Total Households	KS402EW
47	Tenure: Rent (Public)	Housing Type	v16	v32	Percentage	Total Households	KS402EW
48	House: Semi-detached	Housing Type		v26	Percentage	Total Households	KS401EW
49	House: Terraced	Housing Type	v18	v29	Percentage	Total Households	KS401EW
50	Household: 2+ Cars	Socio-economic	v26	v41	Percentage	Total Population (16 to 74)	KS404EW
51	Employment: Agriculture/Fishing	Socio-economic	v35	v48	Percentage	Total Employed Population	KS605EW
52	Employment: Administrative activities	Socio-economic		v58	Percentage	Total Employed Population	KS605EW
53	Employment: Education	Socio-economic		v60	Percentage	Total Employed Population	KS605EW
54	Employment: Financial Intermediation	Socio-economic	v40	v53	Percentage	Total Employed Population	KS605EW
55	Employment: Utilities	Socio-economic		v55	Percentage	Total Employed Population	KS605EW
56	Employment: Hotel / Catering	Socio-economic	v38	v51	Percentage	Total Employed Population	KS605EW
57	Employment: Health / Social work	Socio-economic	v39	v52	Percentage	Total Employed Population	KS605EW
58	Employment: ICT activities	Socio-economic		v57	Percentage	Total Employed Population	KS605EW
59	Employment: Manufacturing	Socio-economic	v37	v50	Percentage	Total Employed Population	KS605EW
60	Employment: Mining/Quarrying/Construction	Socio-economic	v36	v49	Percentage	Total Employed Population	KS605EW
61	Employment: Public administration and defence	Socio-economic		v59	Percentage	Total Employed Population	KS605EW
62	Employment: Transport	Socio-economic		v56	Percentage	Total Employed Population	KS605EW
63	Employment: Wholesale/retail trade	Socio-economic	v41	v54	Percentage	Total Employed Population	KS605EW
64	Economically Inactive: Provide unpaid care	Socio-economic	v30	v36	Percentage	Total Population	KS301EW
65	Economically Inactive: Looking after family	Socio-economic	v34		Percentage	Total Population (16 to 74)	KS601EW
66	Economically Inactive: Full-Time Student	Socio-economic	v31	v40	Percentage	Total Population (16 and over)	KS601EW
67	Economically Active: Working full-time	Socio-economic		v47	Percentage	Total Population (16 to 74)	KS601EW
68	Economically Inactive: limited by long term illness	Socio-economic	v29	v35	Percentage	Total Population	KS601EW
69	Economically Active: Working part-time	Socio-economic	v33	v46	Percentage	Total Population (16 to 74)	KS601EW
70	Occupation: Routine/Semi-Routine	Socio-economic	v25		Percentage	Total Population (16 to 74)	KS611EW
71	Economically Inactive: Unemployed	Socio-economic	v32	v45	Percentage	Total Population (16 to 74)	KS601EW
72	Travel to Work: foot, Bicycle or Other	Socio-economic		v44	Percentage	Total Population (16 to 74)	QS701EW
73	Travel to Work: Work from home	Socio-economic	v28		Percentage	Total Population (16 to 74)	QS701EW
74	Travel to Work: Private Transport	Socio-economic		v43	Percentage	Total Population (16 to 74)	QS701EW
75	Travel to Work: Public Transport	Socio-economic	v27	v42	Percentage	Total Population (16 to 74)	QS701EW
76	Qualification: Level 1, 2 or Apprenticeship	Socio-economic		v37	Percentage	Total Population (16 and over)	QS501EW
77	Qualification: Level 3	Socio-economic		v38	Percentage	Total Population (16 and over)	QS501EW
78	Qualification: Higher Education (L4)	Socio-economic	v24	v39	Percentage	Total Population (16 and over)	QS501EW

B.10 Brainstorming Hierarchical Designs

VARIABLES		SCALE		EXTENT
RESOLUTION		SCALE		EXTENT
<p>VARIABLES</p> <p>Variables as Level 1. If Variables appear as Level 2 or 3 then 1 and 2 need to have limited numbers</p> <p>Matrix global value of CC Transition from GlobalMany = quick starting point: reduce by significant values, re-order by values - max,min, mean etc</p>  <p>Juxtaposition: Space</p>		<p>Multi View:</p> <p>Superposition</p>  <p>Level 1: Variables Level 2: Scale i.e. show 5 spatial aggregations in one cell - show all in matrix</p> <p>Multi View:</p> <p>Juxtaposition</p>  <p>Level 1: 3 matrices for 3 different locations Level 2: 3 within 1 cell: also with overall total:</p> <p>SPATIAL</p> <p>Spatial View:</p> <ul style="list-style-type: none"> - compare series of maps for resolution A with series of maps for resolution B becomes difficult with more V <p>Statistical View</p> <ul style="list-style-type: none"> - show all SR in SPLOM or as multiple SPLOMS  <p>SPLOM for SR 1: SPLOM for SR 2:</p>		<p>Multi View:</p> <p>Juxtaposition</p>  <p>Level 1: 3 matrices for 3 different locations Level 2: 3 within 1 cell: also with overall total:</p> <p>SPATIAL</p> <p>Spatial View:</p> <ul style="list-style-type: none"> - compare series of maps for extent A with series of maps for extent B becomes difficult with more V <p>Statistical View</p> <ul style="list-style-type: none"> - like sampling - extent can be geographical, random, stratified (temporal or attribute based)- scatterplot either colour coded by extent to compare or different scatter
<p>GEOGRAPHY</p> <p>No local stats (global) but can use AR values for choropleth or difference maps</p> <p>Series of maps for each V (not in matrix): choropleth, cartogram or treemaps of the mean value at AR (or matrix of composite maps).</p> <p>Maps can be reordered by mean, max, name etc</p> <p>V becomes many (too many) when no longer possible to fit maps on screen or visually compare them (i.e. less easy if AR is small units)</p>		<p>Attribute</p> <p>Superposition</p>  <p>As above but temporal i.e. 4 seasons = 1 year</p> <p>Attribute</p> <p>Same as Spatial</p> <p>i.e. compare periods of time with the same variables i.e. data for WINTER compared to the same variables for SUMMER</p>		<p>Attribute</p> <p>Same as Spatial</p> <p>i.e. compare periods of time with the same variables i.e. data for WINTER compared to the same variables for SUMMER</p>
<p>TRANSFORMATION</p> <p>Juxtaposition + asymmetrical matrix</p>  <p>Above: Logged values Below: non-logged values</p> <p>Juxtaposition + symmetrical matrix</p>  <p>i.e. Difference value</p> <p>Juxtaposition + Superposition + symmetrical matrix</p>  <p>i.e. Multiple difference values</p>		<p>TEMPORAL</p> <p>Superposition</p>  <p>As above but attributes i.e. two variables combined to be one</p> <p>TEMPORAL</p> <p>Not same as Spatial</p> <p>Juxtaposition - switch from large matrix to smaller Different number of variables so comparison side-by-side not necessary</p>		<p>TEMPORAL</p> <p>Not same as Spatial</p> <p>Juxtaposition - switch from large matrix to smaller Different number of variables so comparison side-by-side not necessary</p>

VARIABLES		RESOLUTION	SCALE	EXTENT
Juxtaposition		within each grid square show aggregate scale		
Usually V will be Level 1: as if Level 2 or 3 limited by space The more V the more difficult it is to show L		Scatter: can also show change in number of areas		Juxtaposition
		map of squares - containing smaller squares		can show correlation maps next to each other: either as maps or as treemap - but as location is different tree map maybe confusing (Only comparing colour /values as same variables)
			SPATIAL	
			Juxtaposition	
			Also in Level 2: but trade off between v & L for detail	
			(Show local variation of scale!)	
			TEMPORAL	
			Superposition + juxtaposition - (map) as above	Same as Spatial
			As above but temporal i.e. 4 seasons = 1 year	i.e. compare periods of time with the same variables i.e. data for WINTER compared to the same variables for SUMMER
			ATTRIBUTE	
			Superposition + Juxtaposition (map) see above	Not same as Spatial
			As above but attributes i.e. two variables combined to be one	Switch from large number of variables to fewer (i.e. large matrix to smaller).
				Different number of variables so comparison side-by-side not necessary

VARIABLES		RESOLUTION	SCALE	EXTENT
Juxtaposition	Distribution of the variable locally - i.e. 1 statistical value (mean, skewness etc - see overview) per local area - either variable distribution can be represented as a single map/statistical view in DoD or by using the blank square of the matrix where Column = Row and thus allowing for comparison of variable distribution across variables.	Superposition  Summary value within each grid square can show aggregate scale	SPATIAL within Juxtaposition map of squares - containing smaller squares which show scale	Juxtaposition  can show extent next to each other: as maps with different boundaries - i.e UK v London comparing colour /values of same variables side by side
GEOGRAPHY Show statistical view of distribution: either Multiple box plot plots or as histograms on a map. But difficult to show either of these in the central matrix line - just DoD view Max L = where too many box plots to visualise		Map: of local statistics (skewness, etc) : either as choropleth, cartogram or treemap  Although Descriptive variables are not part of correlation comparison: Can use distribution maps in a matrix - with DoD view  Level 1b: Variables as Level 1, Map of Extent + local stats in centre cells DoD: Map of Scale:Extent + Local Stats	Superposition within Juxtaposition see above  As above but temporal i.e. 4 seasons = 1 year TEMPORAL i.e. compare periods of time with the same variables i.e. data for WINTER compared to the same variables for SUMMER Same as Spatial	Not same as Spatial Juxtaposition Time - switch from large number of variables to fewer (i.e. large matrix to smaller). ATTRIBUTE Different number of variables so comparison side-by-side not necessary
TRANSFORMATION Compare / Show difference calculation: in spatial or non-spatial view. i.e. diverging legend different scale to show how the log changes the distribution or juxtaposition: show change from one to other (include correlation change)		Superposition within Juxtaposition see above  As above but attributes i.e. two variables combined to be one		

Bibliography

- ADDResponse. Addressponse: Auxiliary data driven nonresponse bias analysis blog, 2014. URL <https://blogs.city.ac.uk/addressponse/>. [Accessed on: 2015-01-14].
- M. Adnan. *Towards real-time geodemographic information systems: design, analysis and evaluation*. PhD thesis, UCL (University College London), 2011. URL <http://discovery.ucl.ac.uk/1335608/>. [Accessed on: 2013-08-01].
- M. Adnan, A. Singleton, and P. Longley. Spatially weighted geodemographics. In *GIS Research UK 21st Annual Conference*, Liverpool University, April 2013.
- A. Al-Awami, J. Beyer, H. Strobel, N. Kasthuri, J.W. Lichtman, H. Pfister, and M. Hadwiger. Neurolines: A subway map metaphor for visualizing nanoscale neuronal connectivity. In *IEEE Transactions on Visualization and Computer Graphics (Proceedings IEEE InfoVis)*, 2014.
- H. Allcott. Social Norms and Energy Conservation. *Journal of Public Economics*, 95(9-10):1082–1095, 2011.
- G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M-J. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *IJGIS*, 24(10):1577–1600, 2010.
- L. Anselin, I. Syabri, and O. Smirnov. Visualizing multivariate spatial correlation with dynamically linked windows. In *University of California, Santa Barbara. CD-ROM*, 2002.
- B. Asare-Bediako, L. M. Ramirez Elizondo, P. F. Ribeiro, W. L. Kling, and G. C. Paap. Consideration of electricity and heat load profiles for intelligent energy management systems. In *Universities' Power Engineering Conference (UPEC), Proceedings of 2011 46th International*, pages 1–6, 2011.
- D. I. Ashby and P. A. Longley. Geocomputation, Geodemographics and Resource Allocation for Local Policing. *Transactions in GIS*, 9(1):53–72, January 2005.
- J. Bertin. *Semiologie graphique: les diagrammes, les rseaux, les cartes*. La Haye, Mouton; Gauthier-Villars, Paris, Paris, 1967.
- B. Boardman. New directions for household energy efficiency: evidence from the UK. *Energy Policy*, 32(17):1921–1933, November 2004.
- H. Bohnacker. *Generative Design: Visualize, Program, and Create with Processing*. Princeton Architectural Press, New York, August 2012.
- D. Bonino, F. Corno, and L. De Russis. Home energy consumption feedback: A user survey. *Energy and Buildings*, 47:383–393, April 2012.
- A. Boucher, D. Cameron, and N. Jarvis. Power to the people: Dynamic energy management through communal cooperation. In *Proceedings of the Designing Interactive Systems Conference, DIS '12*, pages 612–620, New York, NY, USA, 2012. ACM.
- K. Brennan. *A Guide to the Business Analysis Body of Knowledge*. International Institute of Business Analysis, 2nd edition, 2009.
- C. Brunson. Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science*, 39(3):497–524, 1999.
- C. Brunson, A. Fotheringham, S., and M. E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.
- CACI. Acorn technical document. Technical report, CACI Product Development Team, CACI Ltd, Kensington Village, London, May 2014. URL <http://acorn.caci.co.uk/downloads/Acorn-Technical-document.pdf>. [Accessed on: 2014-10-14].

- M. Callingham. Putting geography back into geodemographics. In *GIS Research UK 15th Annual Conference*, National University of Ireland Maynooth, April 2007.
- P. Campanelli, P. Sturgis, and S. Purdon. Can you hear me knocking: An impact into the impact of interviewers on survey response rates. Technical report, The Survey Methods Centre, SCPR, 1997.
- N. Cao, D. Gotz, J. Sun, and H. Qu. DICON: interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2581–2590, December 2011.
- J. Cheshire and O. Uberti. *LONDON: The Information Capital: 100 maps and graphics that will change how you view the city*. Particular Books, October 2014.
- G Chicco, R Napoli, and F Pigliane. Comparisons Among Clustering Techniques for Electricity Customer Classification. *IEEE Transactions on Power Systems*, 21(2):933–940, May 2006.
- J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), IEEE Symposium on*, pages 27–34, October 2010.
- C. Clastres. Smart Grids: Another Step Towards Competition, Energy Security and Climate Change Objectives. *Energy Policy*, 39:5399–5408, 2011.
- S. Cockings, A. Harfoot, D. Martin, and D. Hornby. Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for england and wales. *Environment and Planning A*, 43(10):2399–2418, 2011.
- M. Cohn. Techniques for estimating. In *Agile Estimating and Planning*, pages 49 – 60. Addison-Wesley, Boston, 2005.
- E. Costanza, S. D. Ramchurn, and N. R. Jennings. Understanding domestic energy consumption through interactive visualisation: A field study. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp ’12, pages 216–225, New York, NY, USA, 2012. ACM.
- T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. CRC Press, Boco Raton, Florida, 2nd edition, 2001.
- P. Cruz. Empires Decline: Revisited, 2010. URL <http://bit.ly/1zeoskt>. [Accessed on: 2014-10-28].
- S. Darby. Smart Metering: What Potential for Householder Engagement? *Building Research & Information*, 38(5):442–457, 2010.
- A. de Almeida, P. Fonseca, B. Schlomann, and N. Feilberg. Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations. *Energy and Buildings*, 43:1884–1894, 2011.
- L. De Silva, C. Morikawa, and I. Petra. State of the art of smart homes. *Engineering Applications of Artificial Intelligence*, 25:1313–1321, 2012.
- D. Dean, J. Hender, T. Rodgers, and E. Santanen. Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Assoc. for Information Systems*, 7(10), November 2006.
- DECC. LLSOA Electricity and Gas Consumption Data – Experimental. Technical report, Department of Energy and Climate Change, 2008. URL <http://bit.ly/DeccData08y>. [Accessed on: 2014-10-25].
- DECC. Consumption Statistics: sub-national residual fuel. Technical report, Department of Energy and Climate Change, 2012. URL <http://bit.ly/DeccStats>. [Accessed on: 2014-10-29].
- DECC. MSOA/IGZ and LSOA gas and electricity statistics: Methodology and Guidance. Technical report, Department of Energy and Climate Change, 2013a. URL <http://bit.ly/1c7ghe3>. [Accessed on: 2013-07-25].
- DECC. LSOA estimates of households not connected to the gas network. Technical report, Department of Energy and Climate Change, 2013b. URL <http://bit.ly/NoGasLSOA>. [Accessed on: 2014-10-29].
- DECC. Fuel poverty: a framework for future action. Technical report, Department of Energy and Climate Change, 2013c. URL <http://bit.ly/fuelPoverty>. [Accessed on: 2013-09-08].

BIBLIOGRAPHY

- DECC. Energy consumption in the united kingdom (ECUK): user guide. Technical report, Department of Energy and Climate Change, 2013d. URL <http://bit.ly/UKGovEnergy>. [Accessed on: 2013-08-25].
- DECC. Energy consumption in the UK (2014). Technical report, Department of Energy and Climate Change, July 2014. URL <http://bit.ly/UKEnergy>. [Accessed on: 2014-10-13].
- T. Dillahunty and J. Mankoff. In the Dark, Out in the Cold. *XRDS: Crossroads, The ACM Magazine for Students - Green Technologies*, 17(4):39–41, June 2011.
- T. Dillahunty, J. Mankoff, E. Paulos, and S. Fussell. It’s Not All About ”Green”: Energy Use in Low-Income Communities. In *Proceedings of the 11th international conference on Ubiquitous computing*, Ubicomp, pages 255–264, NY, USA, 2009. ACM.
- D. Dorling. *Area Cartograms: Their Use and Creation, Concepts and Techniques in Modern Cartography (CATMOG, 59)*. University of Bristol, England, 1996.
- G. Dove and S. Jones. Evaluating creativity support in co-design workshops. In *Paper presented at the CHI 2013 Workshop: Evaluation Methods for Creativity Support Environments*, Paris, France, 2013.
- G. Dove and S. Jones. Using data to stimulate creative thinking in the design of new products and services. In *ACM Designing Interactive Systems*, Vancouver, Canada, 2014. ACM Press.
- S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction*, 17(4):18:1–18:24, December 2010.
- A. B. Downey. *Think Stats*. O’Reilly Media Inc, Sebastopol, CA, 2011.
- A. Druckman and T. Jackson. Household Energy Consumption in the UK: A Highly Geographically and Socio-economically Disaggregated Model. *Energy Policy*, 36:3177–3192, August 2008.
- A. Duffy. *Opening the Door to Personal Metaphor Within Organisations: A study of everyday metaphor in the workplace*. MSc. thesis: Innovation, creativity and leadership, City University London, Mar 2013.
- J. Dykes and C. Brunsdon. Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis. *IEEE TVCG*, 13(6):1161–1168, 2007.
- J. Dykes and D. Lloyd. Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study. *Transactions in Visualization and Computer Graphics*, 17(6), 2011.
- J. Dykes, A. M. MacEachren, and M.-J. Kraak. *Exploring Geovisualization: International Cartographic Association*. Pergamon, February 2005.
- J. Dykes, J. Wood, and A. Slingsby. Rethinking map legends with visualization. *IEEE TVCG*, 16(6): 890–899, 2010.
- K. Ehrhardt-Martinez, K. Donnelly, and J. Laitner. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Technical Report Research Report E105, ACEEE, June 2010. URL <http://www.aceee.org/research-report/e105>. [Accessed on: 2011-11-22].
- K. Ellegård and J. Palm. Visualizing energy consumption activities as a tool for making everyday life more sustainable. *Applied Energy*, 88:1920–1926, 2011.
- Energy Saving Trust. Powering the Nation: Household electricity-using habits revealed. Technical report, Energy Saving Trust, London, 2012. URL <http://bit.ly/PowerNation>. [Accessed on: 2014-10-28].
- E.ON Technology & Innovation. Project completion summary: Visualizing energy use to facilitate development and uptake of smart home services, 2014. URL <http://bit.ly/EonComplete>. [Accessed on 2014-10-28].
- Experian. GreenAware on INSOURCE: Segment and target consumers based on their environmental awareness, 2008. URL <http://bit.ly/GreenAware>. [Accessed on: 2014-10-10].
- Experian. Mosaic united kingdom: The consumer classification of the united kingdom. Technical report, Experian Ltd, Nottingham, 2009. URL <http://bit.ly/Mosaic2009>. [Accessed on: 2013-08-21].

- Experian. Mosaic: The consumer classification solution for consistent cross-channel marketing, 2014. URL <http://bit.ly/MosaicExp>. [Accessed on: 2014-10-10].
- Z. Fan, Q. Chen, G. Kalogridis, S. Tan, and D. Kaleshi. The power of data: Data analytics for M2M and smart grid. In *2012 3rd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe)*, pages 1–8, 2012.
- A. Faruqui, D. Harris, and R. Hledik. Unlocking the 53 Billion Euro Savings from Smart Meters in the EU. *Energy Policy*, 38:6222–6231, 2010.
- J-D. Fekete, J. van Wijk, J. T. Stasko, and C. North. The value of information visualization. In A. Kerren, J. T. Stasko, J-D. Fekete, and C. North, editors, *Information Visualization*, Lecture Notes in Computer Science, pages 1–18. Springer Berlin Heidelberg, January 2008.
- S. Few. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press, Oakland, Calif, 1st edition, September 2004.
- S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, Calif, 1st edition, April 2009.
- A. Field. *Discovering Statistics using IBM SPSS Statistics*. SAGE Publications Ltd, Los Angeles, 4th edition, January 2013.
- S. Firth, K. Lomas, A. Wright, and R. Wall. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings*, 40(5):926–936, January 2008.
- P. Fisher, N. Tate, and A. Slingsby. Type-2 fuzzy sets applied to geodemographic classification. In *Eighth International Conference on Geographic Information Science (GIScience 2014)*, Vienna, Austria., September 2014.
- A. S. Fotheringham and D. W. S. Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.
- A. S. Fotheringham, C. Brunson, and M. Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester, England ; Hoboken, NJ, USA, 1st edition, October 2002.
- B. Fry. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O’Reilly Media, Sebastopol, CA, 1st edition, Jan 2008.
- C. G. Gale. Personal Communication, 2014a.
- C. G. Gale. *Creating an Open Geodemographic Classification using the UK Census of the Population*. PhD thesis, UCL (University College London), may 2014b.
- C. G. Gale and P. A. Longley. Geodemographic output area classifications for london, 2001-2011. In *GIS Research UK*, Lancaster University, Lancaster UK., apr 2012.
- C. G. Gale and P. A. Longley. Temporal uncertainty in a small area open geodemographic classification. *Transactions in GIS*, 17:563–588, aug 2013.
- C. G. Gale, M. Adnan, and P.A. Longley. Open Geodemographics: Open Tools and the 2011 OAC. In *GIS Research UK*, Lancaster University, Lancaster UK., April 2012.
- C. G. Gale, P.A. Longley, and A. Singleton. Does london need a separate geodemographic classification? In *8th International Conference on Geographic Information Science*, 2014.
- A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206, July 1992. ISSN 1538-4632.
- V. Giordano and G. Fulli. A business case for Smart Grid technologies: A systemic perspective. *Energy Policy*, 40:252–259, January 2012.
- M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, October 2011.
- S. Goodwin and J. Dykes. Do Local Information Systems Hide the Bigger Picture? An analytical approach to measuring the strength of local boundaries. In *GISRUK*, Manchester Metropolitan University, Manchester UK., April 2008.

BIBLIOGRAPHY

- S. Goodwin, J. Dykes, S. Jones, I. Dillingham, G. Dove, A. Duffy, A. Kachkaev, A. Slingsby, and J. Wood. Creative User-Centered Visualization Design for Energy Analysts and Modelers. *IEEE Transactions on Visualization and Computer Graphics*, 19:2516–2525, 2013.
- W. J. Gordon. *J.(1961) Synectics: The Development of Creative Capacity*. New York: Harper & Row, 1960.
- D. Goulvent. *Household Energy Consumption Classification of Greater London*. MSc. Thesis, Birkbeck University, September 2012. URL <http://dx.doi.org/10.6084/m9.figshare.104658>. [Accessed on: 2013-08-06].
- J. K. Gruber and M. Prodanovic. Residential Energy Load Profile Generation Using a Probabilistic Approach. In *Computer Modeling and Simulation, 6th European Symposium on*, pages 317–322, Valetta, Malta, November 2012.
- S. O. Guerra. Behavioural patterns and user profiles related to energy consumption for heating. *Energy and Buildings*, 43(10):2662–2672, October 2011.
- G. Haq and A. Owen. Green Streets The Neighbourhood Carbon Footprint of York, 2009. URL <http://bit.ly/1wG6bYX>. [Accessed on: 2014-10-29].
- T. Hargreaves, M. Nye, and J. Burgess. Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. *Energy Policy*, 38(10):6111–6119, October 2010.
- T. Hargreaves, M. Nye, and J. Burgess. Keeping energy visible? Exploring how householders interact with feedback from smart energy monitors in the longer term. *Energy Policy*, 52:126–134, January 2013.
- P. Harrington. *Machine Learning in Action*. Manning Publications, NY, 2012.
- P. Harris, C. Brunsdon, and M. Charlton. Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736, 2011.
- P. Harris, C. Brunsdon, M. Charlton, S. Juggins, and A. Clarke. Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences*, 46(1):1–31, January 2014.
- R. Harris, P. Sleight, and Webber R. *Geodemographics: GIS and Neighbourhood Targeting*. Wiley-Blackwell, 2005.
- R. Harris, A. Singleton, D. Grose, C. Brunsdon, and P. Longley. Grid-enabling geographically weighted regression: A case study of participation in higher education in england. *Transactions in GIS*, 14(1): 43–61, 2010.
- L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber’s law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, December 2014.
- M. Harrower and C. A. Brewer. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, June 2003.
- A Hevner and S Ram. Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75–105, March 2004.
- L. Hohmann. *Innovation Games: Creating Breakthrough Products Through Collaborative Play*. Boston: Addison-Wesley, 2007.
- T. G. Holmes. Eco-visualization: combining art and technology to reduce energy consumption. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & Cognition, C&C ’07*, pages 153–162, New York, NY, USA, 2007. ACM.
- S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Moller. DimStiller: Workflows for dimensional analysis and reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 3–10. IEEE, October 2010.
- S. G. Isaksen, K. J. Lauer, and G. Ekvall. Situational outlook questionnaire: A measure of the climate for creativity and change. *Psychological Reports*, 85(2):665–674, 1999.
- S.G. Isaksen, K.B. Dorval, and D.J. Treffinger. *Creative Approaches to Problem Solving: a Framework for Innovation and Change*. SAGE, Los Angeles; London, 3rd edition, 2011.

- J. E. Jackson. *A User's Guide to Principal Components*. John Wiley and Sons, New York, USA, 1991.
- A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3): 264–323, 1999.
- M. G. Jennings. A smarter plan? a policy comparison between great britain and ireland's deployment strategies for rolling out new metering technologies. *Energy Policy*, 57:462–468, June 2013.
- R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson, Upper Saddle River, N.J, 6th edition, April 2007.
- S. Jones, P. Lynch, N. Maiden, and S. Lindstaedt. Use and influence of creative ideas and requirements for a work-integrated learning system. In *16th IEEE International Conference on Requirements Engineering*, pages 289 – 294, Sep 2008.
- L. Jonsson, L. Broms, and C. Katzeff. Watt-Lite: Energy Statistics made Tangible. In *Proceedings of the 8th Conference on Designing Interactive Systems*, DIS, pages 240–243, New York, NY, USA, 2010. ACM.
- D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the information age: Solving problems with visual analytics*. Eurographics Association, Goslar, Germany, 2010.
- L. C. Koh, A. Slingsby, J. Dykes, and T.S. Kam. Developing and Applying a User-Centered Model for the Design and Implementation of Information Visualization Tools. In *InfoVis*, pages 90–95. IEEE, July 2011.
- J. Krause, A. Perer, and E. Bertini. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623, December 2014.
- N. S. Lamand and D. A. Quattrochi. On the issues of scale, resolution and fractal analysis in the mapping sciences. *The Professional Geographer*, 44(1):88–98, 1992.
- J. Ledolter. Clustering. In *Data Mining and Business Analytics with R*, pages 196–219. John Wiley & Sons Inc, Hoboken, New Jersey, 2013.
- G. S. Linoff. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. Wiley, Chichester, 3rd edition, 2011.
- C. D. Lloyd. *Exploring Spatial Scale in Geography*. John Wiley & Sons, Ltd, Chichester, England, 2014.
- P. A. Longley and A. D. Singleton. Classification through consultation: public views of the geography of the e-society. *International Journal of Geographical Information Science*, 23(6):737–763, 2009. [Accessed on: 2013-08-25].
- P. A. Longley, M. Adnan, and G. Lansley. Spatio-temporal demographic classification of the twitter users. In *8th International Conference on Geographic Information Science*, 2014.
- B Lu, P Harris, M Charlton, C Brunson, T Nakaya, and I Gollini. Package GWmodel, 2014. URL <http://cran.r-project.org/web/packages/GWmodel/GWmodel.pdf>. [Accessed: 2014-10-22].
- S. Lühr, G. West, and S. Venkatesh. Recognition of emergent human behaviour in a smart home: A data mining approach. *Pervasive and Mobile Computing*, 3:95–116, 2007.
- J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, April 1986.
- M Maher and D Fisher. Using AI to Evaluate Creative Designs. In *Design Creativity, 2nd Int. Conference on*, Glasgow, UK, September 2012.
- N. Maiden, A. Gizikis, and S. Robertson. Provoking creativity: Imagine what your requirements could be like. *IEEE Software*, 21(5):68–75, 2004.
- N. Maiden, C. Ncube, and S. Robertson. Can requirements be creative? experiences with an enhanced air space management system. In *29th IEEE International Conference on ICSE*, pages 632–641, 2007.
- S. T. March and G. F. Smith. Design and natural science research on information technology. *Decision Support Systems*, 15:251–266, 1995.

BIBLIOGRAPHY

- D. Martin. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, 14(1):90–97, 1989.
- D. Martin. Towards the geographies of the 2001 UK census of population. *Transactions of the Institute of British Geographers*, 25(3):321–332, January 2000.
- D. Martin. Last of the censuses? the future of small area population data. *Transactions of the Institute of British Geographers*, 31(1):6–18, March 2006.
- W. Martin. Effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research*, 15(2):304–308, May 1978.
- D.S. Massey and R. Tourangeau. The nonresponse challenge to surveys and statistics. *Annals of the American Academy of Political and Social Science*, 645:6–27, 2013.
- W. J. Matuszak, L. DiPippo, and Y. L. Sun. CyberSAVE: Situational awareness visualization for cyber security of smart grid systems. In *Proceedings of the Tenth Workshop on Visualization for Cyber Security, VizSec '13*, pages 25–32, New York, NY, USA, 2013. ACM.
- T. May, J. Davey, and J. Kohlhammer. Combining statistical independence testing, visual attribute selection and automated analysis to find relevant attributes for classification. In *2010 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 239–240. IEEE, oct 2010.
- T. May, A Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 111–120, October 2011.
- L. T. McCalley and C. J. H. Midden. Energy conservation through product-integrated feedback: The roles of goal-setting and social orientation. *Journal of Economic Psychology*, 23(5):589–603, October 2002.
- D. McCandless. *Information is Beautiful*. Collins, London, December 2012.
- E. McFadzean. The creativity continuum: Towards a classification of creative problem solving techniques. *Creativity and Innovation Management*, 7(3):131–139, 1998.
- S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design activity framework for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2191–2200, December 2014.
- S. McKenna, M. Meyer, C. Gregg, and S. Gerber. s-corrplot: An interactive scatterplot for exploring correlation. *Journal of Computational and Graphical Statistics*, 2015.
- F. McLoughlin, A. Duffy, and M. Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, 48:240–248, 2012.
- A. G. Melville, M. Graham, and J.B. Kennedy. Combined vs. separate views in matrix-based graph analysis and comparison. In *15th International Conference on Information Visualisation (IV)*, pages 53–58, 2011.
- M. Michalko. *Thinkertoys: A Handbook of Creative-Thinking Techniques*. California: Ten Speed Press, December 2010.
- H. J. Miller and J. Han, editors. *Geographic Data Mining and Knowledge Discovery*. CRC Press, Boca Raton, FL, 2nd edition, May 2009.
- M. Monmonier. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21(1):81–84, 1989.
- M. Monmonier. *How to Lie with Maps*. University Of Chicago Press, Chicago, 2nd edition, May 1996.
- T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, November 2009.
- T. Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, Boca Raton, 1st edition, Nov 2014.
- B. J. Oates. *Researching Information Systems and Computing*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, 1st edition, November 2005.

- ONS. 2001 area classifications, 2005. URL <http://bit.ly/0AC2001>. [Accessed on: 2014-10-28].
- ONS. A Beginner's Guide to UK Geography, 2011. URL <http://bit.ly/1eBLXvD>. [Accessed on: 2015-07-26].
- ONS. Methodology note for the 2011 area classification for output areas, 2014. URL <http://bit.ly/0AC11Meth>. [Accessed on: 2014-10-15].
- S. Openshaw and P. Taylor. *The modifiable unit areal problem*. Norwich:Geobooks, 1984.
- A. F. Osborn. *Applied imagination: Principles and Procedures of Creative Problem-Solving*. New York: Scribner, revised edition, 1957.
- A. Paez, M. Trepanier, and C. Morency. Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards. *Transportation Research Part A: Policy and Practice*, 45(7), August 2011.
- J. Palm and K. Ellegård. Visualizing energy consumption activities as a tool for developing effective policy. *International Journal of Consumer Studies*, 35(2):171–179, 2011.
- J. Palmer and I. Cooper. Great Britain's housing energy fact file. Technical report, Department of Energy and Climate Change, 2011. URL <http://bit.ly/GBEnergy>. [Accessed on: 2013-08-24].
- E. Paradis. Package 'ape', dec 2014. URL <http://cran.r-project.org/web/packages/ape/ape.pdf>. [Accessed on: 2015-01-08].
- L. Pennell and N. Maiden. Creating requirements—techniques and experiences in the policing domain. In *Proceedings of REFS 2003 Workshop*, 2003.
- J. Petersen, M. Gibin, P. Longley, P. Mateos, P. Atkinson, and D. Ashby. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, 13(2):173–192, June 2011.
- M. D. Pratt, P. A. Longley, J. Cheshire, and C.G. Gale. Open data sources for domain specific geodemographics. In *GISRUK*, University of Liverpool, Liverpool UK., April 2013.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL <http://www.R-project.org>.
- T. Räsänen, J. Ruuskanen, and M. Kolehmainen. Reducing Energy Consumption by Using Self-Organizing Maps to Create More Personalized Electricity Use Information. *Applied Energy*, 85(9):830–840, September 2008.
- C. Reimann, P. Filzmoser, R. G. Garrett, and R. Dutter. *Statistical Data Analysis Explained*. John Wiley & Sons Ltd, Chichester, England, 2008.
- J. Rodgers. *Residential resource use feedback: exploring ambient and artistic approaches*. MSc. thesis: School of interactive arts and technology, Simon Fraser University, 2011. URL <https://theses.lib.sfu.ca/thesis/etd6474>. [Accessed on: 2014.10.28].
- J. Rodgers and L. Bartram. Exploring Ambient and Artistic Visualization for Residential Energy Use Feedback. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2489–2497, December 2011.
- S. Rusitschka, K. Eger, and C. Gerdes. Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain. In *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 483–488, 2010.
- P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE TVCG*, 11(4):443–456, 2005.
- M. Sedlmair. Personal Communication, 2014.
- M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, dec 2012.
- M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, December 2013.

BIBLIOGRAPHY

- M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, December 2014.
- L. See and S. Openshaw. Fuzzy geodemographic targeting. In Graham Clarke and Moss Madden, editors, *Regional Science in Business*, Advances in Spatial Science, pages 269–281. Springer Berlin Heidelberg, 2001. ISBN 978-3-642-07518-6.
- R. Semenik, R. Belk, and J. Painter. A Study of Factors Influencing Energy Conservation Behaviour. *Advances in Consumer Research*, 09:306–312, 1982.
- J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, July 2002.
- J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005.
- B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, September 1996.
- B. Shneiderman. Inventing discovery tools: Combining information visualization with data mining. *Information Visualization*, 1st(1):5–12, March 2002.
- A. D. Singleton. Comparing classifications: Some preliminary speculations on an appropriate scale for neighbourhood analysis with reference to geodemographic information systems., 2007. URL <http://www.bartlett.ucl.ac.uk/casa/pdf/paper127.pdf>. [Accessed on: 2013-08-15].
- A. D. Singleton. Building a geodemographic classification using gd, 2012. URL <http://rpubs.com/alexsingleton/gd>. [Accessed on: 2013-08-25].
- A. D. Singleton and P. A. Longley. Creating open source geodemographic classifications for higher education applications. Technical report, Centre for Advanced Spatial Analysis, University College London, May 2008.
- A. D. Singleton and P. A. Longley. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29(3):289–298, July 2009a.
- A. D. Singleton and P. A. Longley. Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Regional Science*, 88(3):643–666, 2009b.
- A. Slingsby, J. Dykes, and J. Wood. Configuring Hierarchical Layouts to Address Research Questions. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):977–984, December 2009.
- A. Slingsby, J. Dykes, and J. Wood. Rectangular hierarchical cartograms for socio-economic data. *Journal of Maps*, 6(1):330–345, 2010a.
- A. Slingsby, J. Dykes, J. Wood, and R. Radburn. OAC Explorer: Interactive Exploration and Comparison of Multivariate Socioeconomic Population Characteristics. In M Haklay, J Morley, and H Rahemtulla, editors, *Proceedings of the GIS Research UK*, pages 167–174, 2010b.
- A. Slingsby, J. Dykes, and J. Wood. Exploring Uncertainty in Geodemographics with Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2545–2554, December 2011.
- A. Slingsby, J. Dykes, J. Wood, and R. Radburn. Designing an exploratory visual interface to the results of citizen surveys. *International Journal of Geographical Information Science*, 28(10):2090–2125, October 2014a. ISSN 1365-8816.
- A. Slingsby, N Tate, and P Fisher. Visualisation of uncertainty in a geodemographic classifier. In *Workshop on Visually-Supported Reasoning with Uncertainty at GIScience 2014*, Vienna, Austria, September 2014b.
- L. H. Son, B. C. Cuong, P. L. Lanzi, and N. T. Thong. A novel intuitionistic fuzzy clustering method for geo-demographic analysis. *Expert Systems with Applications*, 39(10):9848–9859, August 2012.
- P. Stephenson, I. Lungu, M. Paun, I. Silvas, and G. Tupu. Tariff development for consumer groups in internal European electricity markets. In *Electricity Distribution, 2001. Part 1: Contributions. CIRED.*, volume 5, Amsterdam, 2001. IEE.

- C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, December 2014.
- A. Tait. Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011. Technical report, Office of National Statistics, Nov 2012a. [Accessed on: 2013-08-05].
- A. Tait. Changes to output areas and super output areas in england and wales, 2001 to 2011. Technical report, Office of National Statistics, November 2012b. URL <http://bit.ly/GeogChanges>. [Accessed on: 2014-10-10].
- M. Theus and S. Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2008. ISBN 1584885947, 9781584885948.
- J. Thomas and J. Kielman. Challenges for visual analytics. *Information Visualization*, 8(4):309–314, 2009.
- J.J. Thomas and K.A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, January 2006.
- Gateway to Research. Using multi-level multi-source auxiliary data to investigate nonresponse bias in uk general social surveys, 2014. URL <http://gtr.rcuk.ac.uk/project/E53910BF-9131-4F9F-8BC8-ED3DE7C3B720>. Active Research Project funded by the ESRC. Project Reference: ES/L013118/1. Lead Research Organisation: City University London [Accessed on: 2015-01-14].
- W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.
- W. R. Tobler. Thirty five years of computer cartograms ground-truthing geodemographics. *Annals of the Association of American Geographers*, 94(1):58–73, 2004.
- R. C. Tryon. *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards brother, Incorporated, lithoprinters and publishers, 1939.
- E. R. Tufte. *The visual display of quantitative information*, volume 2. Graphics Press Cheshire, CT, 1983.
- E. R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, Conn., May 1991.
- J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, London, Reading, Mass, 1977.
- C. Turkay. *Integrating Computational Tools in Interactive and Visual Methods for Enhancing High-dimensional Data and Cluster Analysis*. PhD thesis, University of Bergen, Norway, 2013.
- C. Turkay, A. Lundervold, A.J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, December 2012.
- C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes. Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Transactions on Visualization and Computer Graphics*, Dec 2014.
- A. Vande Moere and H. Purchase. On the role of design in information visualization. *Information Visualization*, 10(4):356–371, October 2011.
- D. Vickers and J. Pritchard. Visualising the output area classification. *Journal of Maps*, 6(1):410–416, January 2010.
- D. Vickers and P. Rees. Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403, 2007.
- D. Vickers, P. Rees, and M. Birkin. A new classification of UK local authorities using 2001 Census key statistics. Technical report, School of Geography, University of Leeds, 2003.
- D. Vickers, P. Rees, and M. Birkin. Working paper 05/2: Creating the national classification of census output areas: data, methods and results. Technical report, School of Geography, University of Leeds, 2005.
- D. W. Vickers. *Multi-Level Integrated Classifications Based on the 2001 Census*. PhD thesis, University of Leeds, January 2006. URL <http://etheses.whiterose.ac.uk/15/>. [Accessed on: 2014-12-01].

BIBLIOGRAPHY

- R. Walker, A. Slingsby, J. Dykes, Kai Xu, J. Wood, P.H. Nguyen, D. Stephens, B.L.W. Wong, and Yongjun Zheng. An extensible framework for provenance in human terrain visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2139–2148, December 2013.
- C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2013.
- M. Weiss, A. Helfenstein, F. Mattern, and T. Staake. Leveraging smart meter data to recognize home appliances. In *IEEE International Conference on Pervasive Computing and Communications*, Lugano, March 2012. IEEE.
- L. Wilkinson. *The Grammar of Graphics*. Springer New York, New York, NY, 2005.
- P. C. Wong, K. Schneider, P. Mackey, H. Foote, Jr Chin, G., R. Guttromson, and J. Thomas. A novel visualization technique for electric power grid analytics. *IEEE Trans Vis Comput Graph*, 15(3):410–423, June 2009.
- J. Wood. Experiments in bicycle flow animation, 2012. URL <http://bit.ly/10f2jie>. [Accessed on: 2014-10-28].
- J. Wood and J. Dykes. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–1355, November 2008.
- J. Wood, R. Beecham, and J. Dykes. Moving beyond sequential design: Reflections on a rich multi-channel approach to data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2014.
- J. Yang, A. Patro, H. Shiping, N. Mehta, M.O. Ward, and E.A Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *IEEE Symposium on Information Visualization (INFOVIS) 2004*, pages 73–80, 2004.
- Y. G. Yohanis, J.D. Mondol, A. Wright, and B. Norton. Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings*, 40:1053–1059, 2008.
- Y. Zhao. *R and Data Mining – Examples and Case Studies*. Academic Press, Elsevier, S.l., 2012.
- J.-P Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evans. Household electricity survey: A study of domestic electrical product usage. Intertek Report R66141, Department for Environment, Food and Rural Affairs - Department of Energy and Climate Change - Energy Saving Trust, Oxford, 2012. URL <http://bit.ly/defraEnergy>. [Accessed on: 2014-10-14].